ISSN 1747-1524

DCC | Digital Curation Reference Manual

Instalment on
"Scientific Metadata"

http://www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata/

Clive Davenhall
NeSC

October 2011
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Reference Manual Instalment on scientific metadata
Creator	Clive Davenhall (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	**This chapter will discuss the role of metadata in curating and understanding often complex datasets. The intended readership is international and comprises people working in digital curation, who can have a range of backgrounds, so no particular expertise in the sciences is assumed.**
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath.
Contributor	Joy Davidson (editor)
Date	10 October 2011 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Davenhall, C. (2011), " Scientific Metadata", *DCC Digital Curation Manual*, J. Davidson, S. Ross, M. Day (eds), Retrieved <date>, from http://www.dcc.ac.uk/resources/curation-reference-manual/scientific-metadata

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and reuse over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

Digital Curation Reference Manual Editors (2010-2011)

- Joy Davidson, Associate Director, Digital Curation Centre (UK)
- Kevin Ashley, Director, Digital Curation Centre (UK)

Digital Curation Reference Manual Editors (2005-2010)

- Seamus Ross, Director, HATII, University of Glasgow (UK)
- Michael Day, Research Officer, UKOLN, University of Bath (UK)

Digital Curation Reference Manual copy editor

- Florance Kennedy, Administrator, Digital Curation Centre

Peer review board members have included

- Neil Beagrie, JISC/British Library Partnership Manager (UK)
- Georg Büechler, Digital Preservation Specialist, Coordination Agency for the Long-term Preservation of Digital Files (Switzerland)
- Filip Boudrez, Researcher DAVID, City Archives of Antwerp (Belgium)
- Andrew Charlesworth, Senior Research Fellow in IT and Law, University of Bristol (UK)
- Robin L. Dale, Program Manager, RLG Member Programs and Initiatives, Research Libraries Group (USA)
- Wendy Duff, Associate Professor, Faculty of Information Studies, University of Toronto (Canada)
- Peter Dukes, Strategy and Liaison Manager, Infections & Immunity Section, Research Management Group, Medical Research Council (UK)
- Terry Eastwood, Professor, School of Library, Archival and Information Studies, University of British Columbia (Canada)
- Julie Esanu, Program Officer, U.S. National Committee for CODATA, National Academy of Sciences (USA)
- Paul Fiander, Head of BBC Information and Archives, BBC (UK)
- Luigi Fusco, Senior Advisor for Earth Observation Department, European Space Agency (Italy)
- Norman Gray, Researcher, Department of Physics and Astronomy, University of Glasgow
- Hans Hofman, Director, Erpanet; Senior Advisor, Nationaal Archief van Nederland (Netherlands)
- Falk Huettmann, PhD Associate Professor, University of Alaska
- Max Kaiser, Coordinator of Research and Development, Austrian National Library (Austria)
- Gareth Knight, Preservation Officer, CeRch, Kings College London
- Carl Lagoze, Senior Research Associate, Cornell University (USA)
- Nancy McGovern, Associate Director, IRIS Research Department, Cornell University (USA)
- Jen Mitcham, Curatorial Officer, Archaeology Data Service, University of York
- Mary Molinaro, Director, Preservation and Digital Programs, University of Kentucky Libraries
- Reagan Moore, Associate Director, Data-Intensive Computing, San Diego Supercomputer Center (USA)
- Sheila Morrissey, Senior Research Developer, ITHAKA
- Alan Murdock, Head of Records Management Centre, European Investment Bank (Luxembourg)
- Julian Richards, Director, Archaeology Data Service, University of York (UK)
- Donald Sawyer, Interim Head, National Space Science Data Center, NASA/GSFC (USA)
- Jean-Pierre Teil, Head of Constance Program, Archives nationales de France (France)
- Mark Thorley, NERC Data Management Coordinator, Natural Environment Research Council (UK)
- Helen Tibbo, Professor, School of Information and Library Science, University of North Carolina (USA)
- Malcolm Todd, Head of Standards, Digital Records Management, The National Archives (UK)
- Andrew Wilson, Senior Data Policy Advisor, Australian National Data Service
- Erica Yang, STFC Rutherford Appleton Laboratory

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and reuse over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Reference Manual* (formerly the Digital Curation Manual) is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resources/curation-reference-manual).

Digital Curation Reference Manual instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and reusers to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually.

To ensure that the manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The list of current and previous members of the peer review board is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed instalments. Both may be sent to the editors of the *DCC Digital Curation Reference Manual* at info@dcc.ac.uk.

Joy Davidson and Kevin Ashley
Digital Curation Centre

18 April 2011

Biography

Clive Davenhall is currently an Analyst / Developer in the Applications Division, Information Services, University of Edinburgh. Until late in 2010 he was a Senior Research Engineer at the National e-Science Centre, Edinburgh, where latterly he worked on the NanoCMOS Project. For NanoCMOS he contributed to *inter alia* the development of the metadata management system. Earlier projects at the NeSC largely involved work on Web Services. Prior to joining the NeSC in February 2005 he worked at the Institute for Astronomy at the Royal Observatory Edinburgh for many years. There he worked on aspects of the astronomers' Virtual Observatory and before that various aspects of astronomical catalogues, databases and archives.

1 Introduction

This chapter of the DCC *Digital Curation Manual* is an introduction to scientific metadata. Metadata is simply defined as 'data about data'; auxiliary data which describes or annotates a dataset. Scientific metadata, then, are auxiliary data and information about scientific datasets. This chapter will discuss the metadata often associated with scientific datasets and the role of these metadata in curating and understanding these often complex datasets. The intended readership is international and comprises people working in digital curation, who can have a range of backgrounds, so no particular expertise in the sciences is assumed. The examples necessarily mostly come from the disciplines with which I am familiar, astronomy and microprocessor circuit simulation.

The term 'metadata' was first used *circa* 1970 in the context of database management systems to denote the additional information needed to describe data items held in such systems. It was commonly used in this context in the later 1970s and subsequently enjoyed more widespread use with the proliferation of digital archives and collections, and digital data generally, from the 1980s onwards [i]. It is now in widespread use.

'Metadata' can mean different things to different people: the term is used in a variety of contexts and there are many definitions. Similarly, there are many different ways to classify metadata. For example, one classification widely used in digital libraries is associated with the Metadata, Encoding and Transmission Standard (METS)[1] and splits metadata into three categories:

Descriptive: used for the discovery and identification of data items.

Structural: concerned with the location of the files that comprise the data and their relation to each other.

Administrative: management information for data items, including storage format, provenance and access rights.

In the case of scientific data we may add the additional category of supporting the interpretation, understanding and use of the data.

Metadata are now widely used, particularly in data management, bibliographic, taxonomic and geospatial [ii] systems, but also in general system design and construction. Metadata also play a central role in the Semantic Web. Further, the proper annotation of datasets, that is their description by means of suitable metadata, is important for their assessment, preservation, proper use and subsequent reuse. It allows such questions to be asked as: what are these data? How may they be accessed? How may they be used? How may they be reused? Without such information a dataset is ultimately just a collection of bytes whose access and interpretation remain a matter of guesswork. Thus, the accurate and effective annotation of datasets with metadata is a central concern of digital curation. This centrality is reflected in the present manual, which has several chapters on the subject. In addition to the present one there are:

- *Metadata*, Michael Day (2005),
- *Archival Metadata*, Marlene van Ballegooie and Wendy Duff (2006),
- *Preservation Metadata*, Priscilla Caplan (2006),
- *Learning Object Metadata*, Lorna Campbell (2007).

You should consult these chapters as needed. In particular Michael Day's introductory chapter on *Metadata* provides useful background and frames much of the present discussion.
The present chapter is concerned with metadata for scientific datasets; how the nature of

scientific data and the scientific process affect the metadata required to describe scientific data and the curation of such metadata. The next section sets the scene by considering the role of data and metadata in science. Subsequent sections describe scientific metadata, present some examples and discuss the digital curation of scientific metadata. The chapter concludes with a brief discussion. An alternative discussion of the best practice for curating scientific metadata, in the context of the Dryad repository, has been given by Greenberg *et al*. (2009).

2 Data and Metadata in Science

2.1 The nature of the scientific enterprise

Before considering the role of data and metadata in science it is worth briefly reprising the nature of the scientific enterprise. 'Science' in modern usage (at least in English-speaking countries) is both (i) an organised, transparent and repeatable way of obtaining reliable information about the material world and (ii) the corpus of knowledge so obtained.

Attempts to investigate and explain the natural world date at least from Classical Antiquity (notably Aristotle and Theophrastus) and there were significant developments during the Middle Ages (for example, by Robert Grosseteste, Roger Bacon and William of Ockham). However, modern science and the 'scientific method' is usually taken to date from the 'Scientific Revolution' of the seventeenth century and the concomitant development of the scientific method (see, for example, Okasha (2002, pp2-11) for a brief outline and Henry (2002) for a more detailed treatment).

There is no agreement amongst philosophers and historians of science on the details of a single scientific method which is applicable to all disciplines and which has been followed at all times, and in any event it is a moot point how closely individual practice follows any ideal. However, there are a number of methods or 'paradigms' into which most science can be categorised. The most familiar are the traditional paradigms of experimental science and descriptive (or natural) science. To these may be added the more recent innovations of 'simulation' and 'data-intensive' science.

The first paradigm, experimental science, proceeds by observing natural phenomena under controlled conditions. Experiments are performed in which reproducible measurements are made (and recorded) under controlled, documented conditions. Hypotheses are formulated to explain the experimental results. These hypotheses are then used to predict additional, hitherto unobserved effects. Finally new experiments are performed to look for the predicted effects and thus distinguish between alternative hypotheses. This process has a number of underlying assumptions, including that the natural world is repeatable (the same set of causes will lead to the same effect at different times and in different places), that the observed behaviour will follow rules and that these rules can usually be expressed mathematically. It is not axiomatic that these assumptions must apply in the natural world, but in practice it has been found that they do [iii].

The whole scientific process should be open, at least in principle, to permit replication and additional testing by other interested parties. Strictly speaking the above adumbration describes 'pure' or curiosity-driven science. Applied science and technology use similar techniques for material benefit. In practice there may be commercial, legal or (particularly in medicine) ethical reasons to limit openness. Such restrictions tend to be more onerous in the applied than the pure sciences.

The second paradigm, descriptive science, is important in the observational or natural sciences such as geology, meteorology, astronomy and some aspects of biology, economics and the social sciences, where direct experiments are not possible. Instead of performing experiments, a survey or census of some type of object under study is undertaken. The resulting collection is studied to attempt to classify the different types of entity it contains and determine the representative characteristics of these entities. The types of entity identified are

then arranged in a classification scheme, or taxonomy, of related types. A good classification scheme reveals underlying relations between the types whereas a bad one merely groups them according to superficial similarities [iv]. This sort of taxonomic classification is particularly important in the biological sciences (see, for example, Figure 1) but also occurs in the other natural sciences.

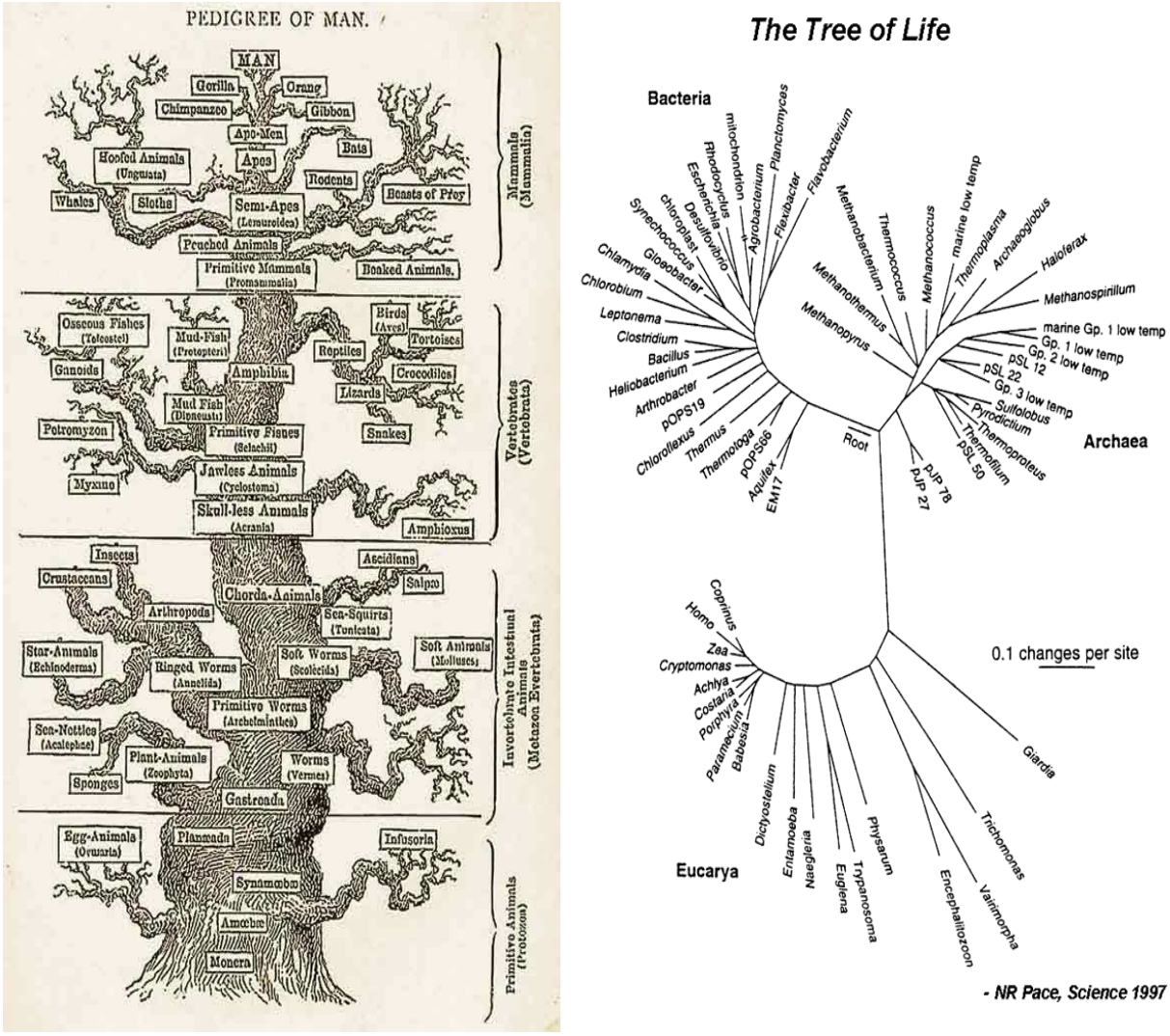


Figure 1: Two representations of the 'tree of life'. Left a classic diagram from Ernst Haeckel's *The Evolution of Man* (1879) and right a modern representation. The tree of life attempts to show the evolutionary (or phylogenetic) relationship between all known forms of life.

In addition to the two traditional components of experiment (or observation) and theory two additional paradigms have been added since the mid-twentieth century: computational simulation and data-intensive science. Both became practical following the development of electronic computers.

Computational simulation is the use of a computer model to simulate complex systems. Traditionally systems are modelled by postulating a theoretical description of the system (for example, using Newtonian gravitation, Maxwell's equations or whatever is appropriate for the system being studied), making suitable assumptions and simplifications, and then finding an analytic solution to the resulting equations describing the system. Many physical systems, particularly natural systems which exist in the real world, are too complex to be studied in this way. Instead, equations describing the system are still established, but they are solved using

numeric or approximate techniques. Complex simulations can require large and complex computer programmes which take a great deal of computer processing time and generate large amounts of data. These data comprise the generated description of the system studied (for an introduction to these techniques see, for example, Shiflet and Shiflet 2006 or, for a more philosophical standpoint, Winsberg 2010).

The final component, data-intensive science, has arisen because both modern experimental and observational science and simulations can generate large amounts of data. Broadly, data-intensive science is concerned with searching and visualising such datasets and the combination of hitherto unrelated datasets to yield new results. These activities have always been part of science, but the scope and scale of modern digital datasets, and development of techniques to manipulate and query them, amount to a new way of doing science: the 'fourth paradigm' discussed by Hey *et al*. (2009) amongst others. The fourth paradigm is closely related to 'data-mining', an ill-defined term concerned with extracting information from large volumes of data, and sometimes the two terms are used interchangeably.

This brief summary of the scientific enterprise will finish with a note on scale. The size of scientific groups or experiments varies enormously. Modern 'big science' projects such as large particle accelerators (notably the Large Hadron Collider, LHC, in Switzerland or the Tevatron in Illinois), orbiting astronomical observatories (such as the Hubble or James Webb Space Telescopes), Earth observation satellites or the Global Biodiversity Information Facility (GBIF) are enormous undertakings costing millions of pounds, dollars or other currency denomination, employing thousands of people (not all of whom are scientists) and lasting for decades. At the other end of the scale individuals or small groups may conduct their own experiments using equipment that they have built themselves. Unsurprisingly there is a corresponding range of practices and procedures between these extremes.

2.1.1 Scientific communication

The scientific method requires successful communication between participants, to report experiments, stimulate new theories and then report and test these new theories. Sufficient detail should be presented to permit replication and further investigation (at least in theory) [v]; this requirement is both good professional practice and common sense.

Books have been produced and circulated, and scholars have communicated by private letters, since Antiquity. However, more modern forms of scholarly communication date from the Scientific Revolution of the seventeenth century and the founding of scientific societies during the same period. These societies organised regular meetings where members could present and discuss results. They also published journals; serial publications with new issues being produced to a regular time-scale, in which results could be reported in greater depth [vi]. While journals became the primary means of presenting new results, books and monographs have continued to have an important role, particularly for consolidating established knowledge and as textbooks for teaching.

Formal publications have never been the whole story. Private communication continues by letter, telephone and more recently email. The circulation of preprints is critically important in many disciplines [vii]. More pertinently, researchers, and to an even greater extent technicians, have always maintained private notes and *aides memoire:* the lab. books and observing logs that contain the actual records of experiments or observations. Further, between private notebooks and formal publications there is a shadowy hinterland of 'grey literature:' internal reports and semi-public documents which may be circulated, and even catalogued and archived, within institutions and groups but which are not formally published. Such material can often contain important information, such as detailed notes and procedures, but can be very difficult to obtain beyond its original circulation. The production of this type of document increased enormously during the second half of the twentieth century [viii]. Initially grey literature was produced on paper though much of it is now electronic.

The traditional means of scholarly communication developed during the print era. Modern electronic means of communication offer an enormous range of additional possibilities which are still being explored. Web sites are an enormously flexible and powerful way of making a large amount of information easily and publicly available. In particular, in the context of scientific communication, they facilitate the production and dissemination of grey literature. Online versions of journals allow traditional publications to be accessed quickly and easily, but more importantly they permit forms of publication that were not feasible hitherto. Particularly, conventional scientific papers presenting results can be linked to the underlying primary data from which the results were generated.

The development of online publication seems likely to have profound consequences for scholarly (and other) publishing, and the full ramifications of these changes are not yet clear. However, they seem likely to affect not just publishing houses but the whole process of scholarly publication. The evolution of scientific communication, with particular reference to the 'fourth paradigm' of data-intensive science is considered by Dirks (2009), Lynch (2009) and other contributors to Part 4 of Hey *et al*. (2009).

2.2 The role of data

As discussed in the previous section, science seeks to obtain reliable knowledge about the natural world by experiment, observation, classification and prediction. The results of experiments are used to identify and categorise the components of the world. Theories are developed to explain experimental results and new experiments performed to test the theories. The whole process should be open, transparent, reproducible and testable.

Data are central to this process. Traditionally data, usually numeric, are generated by experiment or observation. They are reduced and calibrated to a standard or comparable form and perhaps summarised. They are searched to identify and categorise phenomena and used to test predictions and theories.

Often experiments will yield so-called 'raw data;' simple numeric values obtained from the measuring apparatus, perhaps the deflection of a galvanometer needle, adjustment of a vernier or read-out of an analogue-to-digital converter. Such values must be 'reduced' or calibrated into an actual measurement of the physical quantity being studied in some established units, perhaps an electrical potential in volts, magnetic flux density in tesla, or animal densities per hectare, in order to facilitate comparison. The details vary enormously, of course, between experiments, techniques and disciplines, but great care and attention can be required to eliminate insidious systematic effects. In some cases it may be desirable to retain the raw measurements so that the calibration can be improved and the data reduced afresh using the improved values.

Unsurprisingly, given the size and diversity of the scientific enterprise, there is a similar variation in the size, type, complexity and quality of scientific datasets. At one extreme might be a single, one-off experiment, the data from which is reduced and the results published in a regular journal.

At the other extreme are large-scale investigations, surveys, sensor readings and monitoring programmes. Such projects may continue to generate data for years. Monitoring programmes in meteorology and other environmental sciences can continue indefinitely. These undertakings can generate large volumes of data. Further, the data accumulated are often not used to perform a single experiment. Rather, they become a continuing resource, a so-called 'data archive', which can be used and reused in a variety of different investigations. Moreover, the datasets are often sufficiently large (and perhaps still accumulating) that it is impractical and unwieldy to publish them conventionally, for example in a print journal.

Such data archives are not a new phenomenon. For example, the Swan Upping, an annual survey of the swans on the River Thames, has been conducted since the twelfth century [2]. Meteorological observations have been recorded systematically, at least in Western Europe, since the eighteenth century (Golinski 2007). Most observational sciences have similarly acquired archives, which are often housed by museums, government institutions, scientific societies or universities. Originally measurements would have been recorded manually. Later there was automatic recording in the form of strip-charts, photographs and other recording media. Museums also constitute a similar type of resource (see Figure 2). Researchers visit and examine museum collections, photographic libraries and other forms of analogue archive. However, to be useful the collection must be summarised, indexed and catalogued to allow the visitor to find and interpret items of interest.

While data archives are not new, what is different is that most archives are now 'born digital': the data are initially generated in a computer-readable form. This change has been accompanied by an enormous increase in both the total data volume and the rate at which data are generated. As one example, in the nineteenth century major astronomical catalogues, notably the various *Durchmusterungen*, typically listed up to half a million stars, took years to compile and remained in use for the best part of a century. The sky surveys that are their modern equivalents still take over a decade to complete (though there are more of them than there used to be) but they list of the order of a hundred thousand million (10^{11}) stars and galaxies and generate Pbytes of data. However, even these datasets are dwarfed by the datasets generated by modern particle physics experiments or Earth observation programmes. Computer simulations can also generate copiously large amounts of data. All these datasets must be catalogued, indexed and annotated with auxiliary measurements and explanatory information in order to remain useful to their often international users. This situation is not going to change: more, larger and more complex datasets are anticipated for the foreseeable future. There is, for example, now an academic journal devoted solely to databases in the biological sciences: the *Journal of Biological Databases and Curation* [3].

Figure 2: Two examples of physical archives. Left: the Meteorite Hall at the Natural History Museum (Naturhistorisches Museum), Vienna (photograph courtesy Dr Franz Brandstätter and © the Natural History Museum, Vienna). Right: the photographic Plate Library at the Royal Observatory Edinburgh.

Another relatively recent development is the combined archive. Here results from multiple archives or experiments are combined into a single archive, either to allow new results to be

derived (by combining archives of different but related measurements) and/or to allow 'best' values to be derived (by combining different measurements of comparable quantities). The combined archive then becomes a new resource which can be queried and used in subsequent investigations. Considerable, and continuing, effort is required to maintain a combined archive. Particularly, care must be taken in ingesting new results in order to maintain quality control and to ensure the compatibility of data from diverse sources. A relatively early discussion of such combined archives is given by Jaschek (1989). One notable and successful example is the RCSB (Research Collaboratory for Structural Bioinformatics) Protein Data Bank [4]. From a beginning in 1971 when it contained just seven structures, it has now become the single global repository for information on three-dimensional structure of large biological molecules such as proteins and nucleic acids.

Finally, the emerging 'fourth paradigm' of data-driven science is even more ambitious than combined archives. It involves, *inter alia*, searching and combining separate archives and integration with journal literature (Atkinson 2009, Hey *et al*. 2009).

2.3 The role of metadata

Data are central to the scientific enterprise, as outlined in the preceding section. Moreover, there are now many online scientific databases, archives of experimental data, derived archives of combined results and similar archives generated from simulations. In addition the contents of academic journals are often available online and increasingly they can be linked to the appropriate archives in order to access the underlying data from which the reported results were derived. All these resources can be searched to identify results of interest and suitable data extracted, synthesised and reused. More innovatively, facilities adopting the 'fourth paradigm' approach allow disparate resources to be linked and combined in ways not anticipated in advance.

Where are the metadata in this panoply of resources and archives? What role do they play? The answers are that metadata are everywhere and their role is central.

Datasets in archives are usually annotated with metadata. These metadata can be used for a variety of purposes and they are central to the operation of digital archives and the curation of digital data. Data archives vary enormously, both within disciplines as well as across disciplines. The following examples, while by no means exhaustive, suggest some of the common ways in which metadata are used.

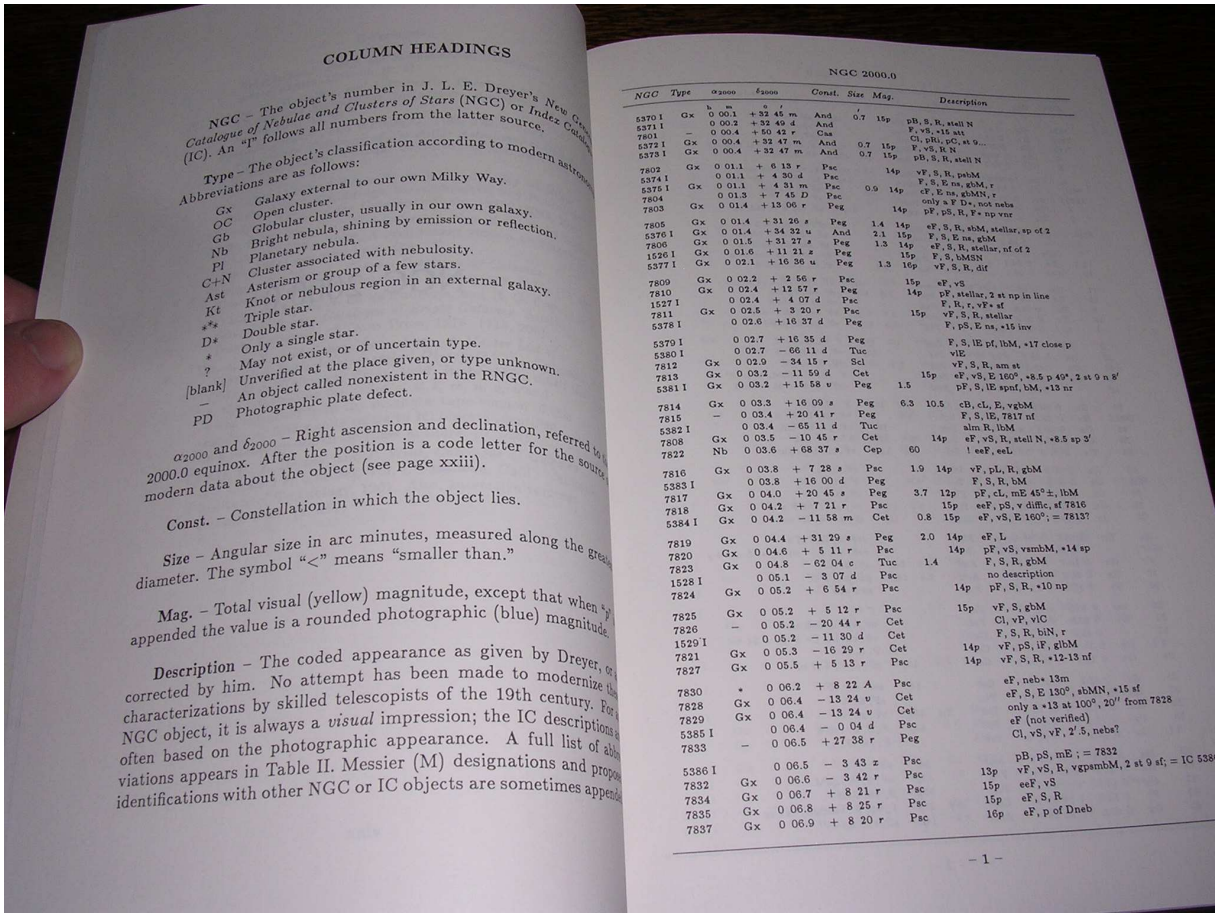
Identify archives that meet a given criteria: metadata comprising summary information about the whole archive (such as the discipline it refers to, the details of the experiment or survey from which it was generated, if appropriate the time and place of the experiment or survey *etc*.) are searched, perhaps as part of a search of a set of archives, to establish whether the archive will be useful for a particular inquiry. Metadata intended to be used in this way are sometimes called 'discovery metadata.'

Provide auxiliary information needed to interpret or extract data: calibration or similar information necessary for the scientific interpretation and analysis of the data in the archive can be stored as metadata. These metadata can be extracted along with the data, thus ensuring that the data are scientifically useful.

Tracking the provenance and processing history of datasets: metadata can record the details of the experiment or survey that generated the data, the nature of any calibration or manipulation applied to the original experimental measurements before they were added to the archive and the origin of external calibration constants or auxiliary data that have been combined with the experimental data. All these details can be important to investigators wishing to use the data.

Metadata are central to the operation of digital archives and the curation of digital data. Though the term metadata only dates from the 1970s the idea of auxiliary data to describe data and annotate datasets is much older. In their simplest form metadata are just book-keeping for data collections. As an arbitrary example, consider Figure 2. The Vienna Museum, like other museums, will maintain a catalogue of its collection of meteorites, as it will for its other collections. For each meteorite this catalogue will include where and when it fell, its classification and how it was acquired by the museum *etc*. Indeed in the photograph the label of each meteorite shows some of this information: metadata on display. Also in Figure 2, in the ROE (Royal Observatory Edinburgh) plate library each plate has its own serial number, which is written on the envelope in which the plate is stored. The ROE maintains a computerised plate catalogue with an entry for each plate. The details for each plate include the area of sky photographed, the length of exposure, the type of emulsion, the date of exposure *etc*. Additional details identify the telescope used to make the exposure and its geographical latitude and longitude.

Star catalogues constitute another example. They are a long-established form of publishing summaries of astronomical results (the earliest catalogue dates from the second century BC and the modern form from at least the nineteenth century). Shorter ones are published in journals, longer ones as books. The basic catalogue comprises a list of stars or other types of object (see Figure 3). The information for each object includes its identifying name or sequence number, celestial coordinates and other measurements summarising its physical properties, all printed in a concise tabular format. Typically at the front of the catalogue there will be a section describing the nature of the catalogue, the procedure used to compile it *etc*. This introduction will be followed by a second section describing the information tabulated. The contents of each column will be described, the units specified and any additional information given. Though the term metadata would never have been used for these details that is exactly what they are: they provide the explanatory information necessary to understand and make use of the data.



COLUMN HEADINGS

NGC – The object's number in J. L. E. Dreyer's *New General Catalogue of Nebulae and Clusters of Stars* (NGC) or *Index Catalogue* (IC). An "I" follows all numbers from the latter source.

Type – The object's classification according to modern astronomical usage. Abbreviations are as follows:

Gx	Galaxy external to our own Milky Way.
OC	Open cluster.
Gb	Globular cluster, usually in our own galaxy.
Nb	Bright nebula, shining by emission or reflection.
Pl	Planetary nebula.
C+N	Cluster associated with nebulosity.
Ast	Asterism or group of a few stars.
Kt	Knot or nebulous region in an external galaxy.
*****	Triple star.
D*	Double star.
*	Only a single star.
?	May not exist, or of uncertain type.
[blank]	Unverified at the place given, or type unknown.
–	An object called nonexistent in the RNGC.
PD	Photographic plate defect.

α_{2000} and δ_{2000} – Right ascension and declination, referred to the *2000.0* equinox. After the position is a code letter for the source of modern data about the object (see page xxiii).

Const. – Constellation in which the object lies.

Size – Angular size in arc minutes, measured along the greatest diameter. The symbol "<" means "smaller than."

Mag. – Total visual (yellow) magnitude, except that when "pg" is appended the value is a rounded photographic (blue) magnitude.

Description – The coded appearance as given by Dreyer, or corrected by him. No attempt has been made to modernize these characterizations by skilled telescopists of the 19th century. For a NGC object, it is always a *visual* impression; the IC descriptions are often based on the photographic appearance. A full list of abbreviations appears in Table II. Messier (M) designations and proposed identifications with other NGC or IC objects are sometimes appended.

NGC 2000.0

NGC	Type	α_{2000} h m	δ_{2000} ° ′	Const.	Size ′	Mag.	Description
5370 I	Gx	0 00.1	+32 45 m	And	0.7	15p	pB, S, R, stell N
5371 I		0 00.2	+32 49 d	And			F, vS, +15 stt
7801	–	0 00.4	+50 42 r	Cas			Cl, pRi, pC, st 9...
5372 I	Gx	0 00.4	+32 47 m	And	0.7	15p	F, vS, R N
5373 I	Gx	0 00.4	+32 47 m	And	0.7	15p	pB, S, R, stell N
7802	Gx	0 01.1	+ 6 13 r	Psc		14p	vF, S, R, psbM
5374 I		0 01.1	+ 4 30 d	Psc			F, S, E ns, gbM, r
5375 I	Gx	0 01.1	+ 4 31 m	Psc	0.9	14p	cF, E ns, glbMN, r
7804		0 01.3	+ 7 45 D	Psc			only a F D's, not nebs
7803	Gx	0 01.4	+13 06 r	Peg		14p	pF, pS, R, F+ np vnr
7805	Gx	0 01.4	+31 26 s	Peg	1.4	14p	eF, S, R, sbM, stellar, sp of 2
5376 I	Gx	0 01.4	+34 32 u	And	2.1	15p	F, S, E ns, gbM
7806	Gx	0 01.5	+31 27 s	Peg	1.3	14p	eF, S, R, stellar, nf of 2
5526 I	Gx	0 01.6	+11 21 s	Peg		15p	F, S, bMSN
5377 I	Gx	0 02.1	+16 36 u	Peg	1.3	16p	vF, S, R, dif
7809	Gx	0 02.2	+ 2 56 r	Psc		15p	eF, vS
7810	Gx	0 02.4	+12 57 r	Peg		14p	pF, stellar, 2 st np in line
1527 I		0 02.4	+ 4 07 d	Psc			F, R, r, vF+ sf
7811	Gx	0 02.5	+ 3 20 r	Psc		15p	vF, S, R, stellar
5378 I		0 02.6	+16 37 d	Peg			F, pS, E ns, +15 inv
5379 I		0 02.7	+16 35 d	Peg			F, S, lE pf, lbM, +17 close p
5380 I		0 02.7	– 66 11 d	Tuc			vlE
7812	Gx	0 02.9	– 34 15 r	Scl			vF, S, R, am st
7813	Gx	0 03.2	– 11 59 d	Cet		15p	eF, vS, E 160°, +8.5 p 49", 2 st 9 n 8'
5381 I	Gx	0 03.2	+15 58 u	Peg	1.5		pF, S, lE spnf, bM, +13 nr
7814	Gx	0 03.3	+16 09 s	Peg	6.3	10.5	cB, cL, E, vgbM
7815	–	0 03.4	+20 41 r	Peg			F, S, lE, 7817 nf
5382 I		0 03.4	– 65 11 d	Tuc			alm R, lbM
7808	Gx	0 03.5	– 10 45 r	Cet		14p	eF, vS, R, stell N, +8.5 sp 3'
7822	Nb	0 03.6	+68 37 s	Cep	60		! eeF, eeL
7816	Gx	0 03.8	+ 7 28 s	Psc	1.9	14p	vF, pL, R, gbM
5383 I		0 03.8	+16 00 d	Peg			F, S, R, bM
7817	Gx	0 04.0	+20 45 s	Peg	3.7	12p	pF, cL, mE 45°±, lbM
7818	Gx	0 04.2	+ 7 21 r	Psc		15p	eeF, pS, v diffic, sf 7816
5384 I	Gx	0 04.2	– 11 58 m	Cet	0.8	15p	eF, vS, E 160°; = 7813?
7819	Gx	0 04.4	+31 29 s	Peg	2.0	14p	eF, L
7820	Gx	0 04.6	+ 5 11 r	Psc		14p	pF, vS, vsmbM, +14 sp
7823	Gx	0 04.8	– 62 04 c	Tuc	1.4		F, S, R, gbM
1528 I		0 05.1	– 3 07 d	Psc			no description
7824	Gx	0 05.2	+ 6 54 r	Psc		14p	pF, S, R, +10 np
7825	Gx	0 05.2	+ 5 12 r	Psc		15p	vF, S, gbM
7826	–	0 05.2	– 20 44 r	Cet			Cl, vP, vlC
1529 I		0 05.2	– 11 30 d	Cet			F, S, R, biN, r
7821	Gx	0 05.3	– 16 29 r	Cet		14p	vF, pS, lF, glbM
7827	Gx	0 05.5	+ 5 13 r	Psc		14p	vF, S, R, +12-13 nf
7830	*	0 06.2	+ 8 22 A	Psc			eF, neb+ 13m
7828	Gx	0 06.4	– 13 24 u	Cet			eF, S, E 130°, sbMN, +15 sf
7829	Gx	0 06.4	– 13 24 u	Cet			only a +13 at 100°, 20" from 7828
5385 I		0 06.4	– 0 04 d	Psc			eF (not verified)
7833	–	0 06.5	+27 38 r	Peg			Cl, vS, vF, 2'.5, nebs?
5386 I		0 06.5	– 3 43 z	Psc			pB, pS, mE ; = 7832
7832	Gx	0 06.6	– 3 42 r	Psc		13p	vF, vS, R, vgpsmbM, 2 st 9 sf; = IC 538
7834	Gx	0 06.7	+ 8 21 r	Psc		15p	eeF, vS
7835	Gx	0 06.8	+ 8 25 r	Psc		15p	eF, S, R
7837	Gx	0 06.9	+ 8 20 r	Psc		16p	eF, p of Dneb

– 1 –

Figure 3: Data and metadata in a traditional, printed star catalogue. On the right is the first page of tabular data and on the left the 'metadata' explaining the columns tabulated. The catalogue shown is *NGC 2000.0* by R.W. Sinnott (1988).

In astronomy at least, the adoption of current ideas about digital scientific data archives was spured in part by the space program. Scientific satellites and space probes, be they for Earth observation, planetary exploration, astronomy, solar physics *etc*, usually return copious amounts of data. They are extremely expensive and any given mission is unlikely to be repeated. It is impossible to inspect or modify the instruments after launch. All these factors are incentives to ensure that the data returned are archived fully and carefully, and all the auxiliary documentation, information and, in the present context, metadata, are present, so that the data can be understood and used. Thus archives of data returned from satellites are often more carefully archived and annotated than those of data from terrestrial sources, but they are not different in principle. The general (and unsurprising) rule of thumb is that the more expensive an experiment or facility, the more likely archiving the data it generates is to be taken seriously.

The widespread, indeed ubiquitous, availability of fast and reliable computer networks has affected the way that data archives are perceived and used. It is now common to access archives remotely, searching them to identify suitable items and then retrieving copies of the selected datasets. Many users no longer work in close proximity to either the archive or its creators. They can be based anywhere in the world and often work in isolation, which further increases the importance of both the appropriate annotation of datasets with metadata and adequate documentation in order to enable the data to be interpreted correctly.

2.4 Scientific software

It is worth making a few remarks about the nature of scientific software. Generalisations should be treated with caution because of the diversity of scientific projects, from single-person efforts to large multi-national collaborations. However, most scientific software is performing specialised, bespoke and often esoteric calculations. The user community, that is the group of people involved in the experiments and data analysis, is small, specialised and knowledgeable. For even the largest of scientific disciplines the number of users is insignificantly small compared to the user base of common PC office or home software [ix]. Moreover, much of this software is developed by the scientists themselves who are not professional programmers and who are working under time and budgetary constraints.

The software produced often reflects the circumstances of its production and is idiosyncratic, difficult to use, poorly documented, of limited applicability, contains hidden or implicit assumptions and produces results in bespoke (and occasionally arcane) formats. The idiosyncratic nature of much scientific software is a well-known but little-discussed issue. Recent informal articles by Love (2009) and Nuin (2010) make some very pertinent points.

Much of the data curated in archives has been processed with such software. In order to track the provenance of the data its metadata should at least identify the software used in the processing. In some cases it may be desirable to retain copies of the software to permit reproducibility. Little work has been done on the curation of the scientific software used to process the data stored in archives. Goble and De Roure (2008) discuss some of the issues, but specifically in the context of work flows and Web services.

In addition the data in archives may be periodically re-reduced as new calibration measurements become available, new calibration techniques are developed and bugs in the processing software are fixed. In such cases it may be necessary to impose version control on the software and include details of the version used in the archive.

2.5 Scientific data formats

Many scientific programmes read and write data in their own *ad hoc*, not to say idiosyncratic, formats. In addition there are also many specialised formats that enjoy limited use within a set of applications in some discipline or sub-discipline. Ilana Stern used to maintain a useful

Scientific Data Format Information FAQ which listed, and provided information about, many of these formats. It is still available [5], but unfortunately has not been updated since 1995. Wikipedia's list of file formats [6] currently lists 20 non-biological scientific data formats and 37 biological ones, and others undoubtedly exist. Examples of a few of the more widely used formats will briefly be mentioned here.

In the physical sciences the more general formats can often represent complex data sets comprising a hierarchical collection of multi-dimensional arrays. They usually also have the ability to include metadata annotating the dataset. A notable early example is the Hierarchical Data System (HDS; Warren-Smith and Lawden 1999) from the Starlink project, which was introduced in the early 1980s and remains in use. Two examples currently in widespread use are HDF and NetCDF. HDF (Hierarchical Data Format) [7] was originally developed at the National Center for Supercomputing Applications (NCSA) and is now supported by the HDF Group. An extension of HDS for Earth observation data, HDF-EOS [8] is available. NetCDF (Network Common Data Form) [9] is largely overseen by the Unidata program at the University Corporation for Atmospheric Research (UCAR). Both are open standards and free or public-domain software to access them is widely available. Both are primarily used in the atmospheric and some geographic sciences.

An example of a rather different sort of format is provided by DICOM (Digital Imaging and Communications in Medicine) [10] which is used in medical imaging. DICOM is more than just a file format and includes, for example, a network communications protocol. However, a file format is part of the standard. A file in this format can contain a set of datasets. Each dataset comprises a set of named attributes. For each dataset one (and only one) attribute holds the image, and the other attributes contain auxiliary information: metadata annotating the image. This approach is adopted to ensure that the image and its associated metadata do not become separated. DICOM was first released in 1985 and it has since been continuously developed. It is overseen by NEMA (National Electrical Manufacturers Association) [11]. DICOM is also known as NEMA Standard PS3, and there is an equivalent ISO standard, 12052. Much more information on medical image formats is available in the Medical Image Format FAQ [12].

The existence of standard data formats such as HDS, HDF, NetCDF, DICOM *etc*. makes it possible to develop general-purpose software to process and display data in one (or more) of the formats. Several such applications will usually be available for any well-established scientific data format.

3 Scientific Metadata

Scientific metadata, as discussed above, documents and annotates scientific datasets, providing the auxiliary information necessary to find, interpret, understand, assess and use them. This section makes a few general remarks about the characteristics and representation of scientific metadata and then introduces a few of the more widely-used scientific metadata formats which may be encountered.

As is probably apparent from the foregoing discussion, there is no clear-cut, simple, distinction between the scientific metadata for a scientific dataset and some 'non-scientific' metadata (such as preservation metadata) which the dataset may acquire. Nor is there any straightforward and unambiguous differentiation between scientific data and the scientific metadata annotating them. Most of the scientific data formats mentioned in Section 2.5 above can incorporate metadata items. In a sense the data and metadata are all data.

3.1 Characteristics and representation

This section will discuss the typical characteristics of scientific metadata. Though some characteristics are common, scientific metadata are as diverse as the scientific data they annotate. Various types of metadata may be encountered, including: natural language words, phrases or longer descriptions intended for human interpretation, tags (typically in the context of XML; see Section 3.1.1, below) and more complex items such as vectors, BLOBS (Binary Large Objects) or complex hierarchical items. However, a common basic metadata item is the 'name value pair' in which some item of information is identified by a name and is given a specified value. The name is, in effect, part of a controlled namespace, but the domain in which this name will be understood can vary and may be ill-defined. The name may be part of the definition of whatever metadata format is being used, in which case the meaning is fixed for all users of the format. Conversely the format may allow users to define their own names and assign meaning to them, in which case the name will only be interpretable amongst whichever group of users has agreed the definition. The diversity and flexibility of scientific datasets means that such 'user defined names' are the norm. Similarly, depending on the format, names may occupy either a flat or a hierarchical name space; the latter, of course, affords better possibilities for avoiding name clashes between different groups of users.

The value specified for the value part of each name / value pair may be either free text, numeric or chosen from a controlled vocabulary where each valid option has a defined meaning. Again, the controlled vocabulary may be part of the data format and hence have universal meaning within the format or be a local convention that only has significance amongst co-operating users. Obviously, for appropriate items, the values may include conventional bibliographic references, URLs or other pointers to additional information or documentation.

Human readable documentation, describing either the data and/or the metadata may also be considered part of the metadata (and certainly part of the information to be curated) though it cannot, in general, be automatically interpreted by machine.

Some of the 'name / value' metadata items will be purely descriptive or qualitative. Others will describe details of the digital representation of the data (format, size, location, *etc*). However, the bulk of the scientific metadata will be auxiliary information necessary to interpret the data. Many of these items will represent physical quantities whose definitions and units must also be recorded.

The values of physical quantities are meaningless unless their units are known. For example, if a velocity is simply specified as the unaccompanied value '37' it could be miles per hour, miles per second, kilometres per hour, kilometres per second or some other measure [x]. The sciences generally use units based on the SI (Système international d'unités) system (BIPM 2006) [13], which has largely replaced the earlier CGS (centimetre, gramme, second) system; both of which are based on the metric system. However, it is important to appreciate that many disciplines and sub-disciplines have their own conventions and arrangements. For example, in optical astronomy, wavelengths are still often measured in Ångström (10^{-10} metre, which is not an SI unit), angles are measured in degrees and hours (the SI system uses radians) with sexagesimal subdivisions into minutes and seconds (a practice ultimately deriving from the number-systems of ancient Mesopotamia and often, but not invariably, also encountered in terrestrial latitudes and longitudes) and brightness in 'magnitudes' (a relative, logarithmic measure that originated in Greek Antiquity and is based on the successive visibility of fainter stars during the onset of twilight).

The measurement of time is another apparently straightforward matter that can cause complications. The differences between calendrical systems that can cause confusion for historians are rarely a problem. However, it is still necessary to be aware of time zones and daylight saving schemes. Most countries use a 24 hour clock in which hours are numbered from 1 to 24, counting from midnight. However, a handful of mostly English-speaking countries number the hours after midnight 1 to 12 and those after noon from 1 to 12 again. In these countries a 24 hour clock may be variously referred to as 'military time', 'astronomical

time', 'railway time' or 'continental time.' For very precise work there are a number of different time systems in existence and it may be necessary to record which is being used. There is an international standard, ISO 8601 (ISO 2004) [14] for representing dates and times which is in widespread use. RFC 3339 (Klyne and Newman 2002), developed by the IETF (Internet Engineering Task Force) [15] is a standard for representing time-stamps on the Internet. It is based on ISO 8601 and though intended as an Internet standard it may be more widely useful for representing computerised dates and times.

There are various ways of representing or 'encoding' metadata name / value pairs. They may be used as part of the format used to store the data, so the data and metadata are stored together, typically in a single file (an example is the FITS format discussed in Section 4.2 below). Alternatively, the metadata may be stored separately from the data, perhaps in a DBMS (as in the NanoCMOS example in Section 4.3 below). A separate store of metadata resources which is intended to support the autonomous inter-operation of distributed resources (as found, for example, in 'fourth paradigm' systems or the Semantic Web) is often referred to as a 'registry'.

There is a standard for metadata registries: ISO/IEC 11179 [16] . It is explicitly intended to support metadata-driven data exchange in an heterogeneous environment. It prescribes which metadata must be stored in the registry, but not the encoding used to represent them. The metadata would typically be encoded using XML or one of the other schemes mentioned below. ISO/IEC 11179 has not been widely adopted, though it has found some use amongst government agencies, particularly in the US.

Various encoding schemes have been used to represent metadata. Sometimes schemes developed for other purposes have been adapted. Examples include MARC (Machine Readable Cataloguing; a scheme intended for representing bibliographic information) [xi], MIME (usually used in electronic mail and the Web protocols) [xii] and its putative replacement DIME [xiii]. However, the ubiquitous XML is now a common way of encoding metadata. Some longer-established metadata standards which were originally defined in terms of SGML have now usually been replaced or complemented with more modern XML versions. Another representation which may be encountered is 'Fielded Text'. Fielded Text and XML are briefly described below.

3.1.1 XML

Since its introduction in 1998 XML (eXtensible Markup Language) has become ubiquitous for representing data on the Web and it is also in widespread use in other contexts. XML is a general format suitable for representing both text documents and arbitrary datasets in an electronic form. This chapter is not the place to describe XML, but many good introductions and tutorials are available. The standard is superintended by the W3C (World-Wide Web Consortium) [17] and the XML home page [18] is a good place to start looking for further information. However, briefly XML is a tag-based language derived from SGML (Standardised General Markup Language) and originally intended for representing documents on the Web. SGML is a well-established standard, developed in the mid-1980s by the publishing industry for representing electronic texts. The familiar Web-page mark-up language HTML is an earlier derivative from SGML. A number of related technologies, such as XSLT (eXtensible Stylesheet Language Transformations) and Xquery are available for transforming and querying XML documents.

XML has proved a convenient way of describing metadata, both scientific and otherwise, and often crops up in this and other contexts in digital curation. It is used in a variety of ways. Examples include simply using XML as a convenient and familiar way of representing metadata and in some cases data. Techniques such as XSLT can be used to transform XML metadata into a different representation. Alternatively, when ingesting data into an archive XML may be a convenient way to represent the metadata, which can then be bundled with their associated data.

3.1.2 Fielded text

Fielded Text [19] is a proposed standard which provides annotating metadata for text files containing tables of values. It can handle the common CSV or 'comma separated value' format or many similar ones. In such files the tabular data are simply entered into a text file with one row of the table per line (or record) and the columns in each row separated by a comma (',').

CSV and similar files are a common way of transferring tabular data because they are easy to understand and create, and can be examined without any special software: all that is required is a text editor (or on Unix systems basic commands such as 'cat' or 'more'; on Windows utilities such as Notepad can be used). Most spreadsheets and databases can import and export tables in CSV format. The problem with it is that the information (or metadata) describing the columns is not included and so often ends up being hard-coded into programmes that access the files.

The Fielded Text proposal allows the column details (that is, the metadata describing the columns) to be specified in a separate file, the so-called 'meta file'. This file is itself in XML format. This approach allows the data in the CSV file to be accessed via the information in the meta file, using the techniques of relational database management systems. Specifically, columns can be referred to by name rather than hard-coded position. This change allows an enormous increase in the flexibility of programmes that access the files.

Though Fielded Text is a new initiative the underlying technique is not new. It was used successfully in, *inter alia*, the astronomical catalogue handling systems Haggis (Davenhall *et al*. 1984) and SCAR (Davenhall 1991) in the 1980s and CURSA (Davenhall *et al*. 2001, Davenhall 2001) in the 1990s.

3.2 Standards for scientific metadata

Many standards for representing scientific metadata have been developed within disciplines, sub-disciplines or individual projects or experiments. Most have achieved only limited and local usage. However, a few of the more widely-used ones, which you might encounter, are summarised in Table 1 and are briefly introduced below. Geospatial standards are discussed separately afterwards in Section 3.3. As mentioned above, the use of XML to represent scientific metadata is now very common; all the standards in Table 1 are either defined in terms of XML or now have an XML schema as an optional way of representing them. DDI and TEI are XML schemes for marking-up structured documents and allowing them to be processed flexibly, which is just the traditional use of XML similar to DOCBOOK or other standard XML mark-up. The remaining systems are more conventional metadata.

Though not strictly a standard, a similar system is the Scientific Metadata Model (Sufi and Mathews 2004) developed by the e-Science Data Management Group (DMG) [20] of the UK Science and Technology Facilities Council (STFC). This model provides a general framework for annotating scientific metadata and is applicable to a wide range of disciplines. It is used for most of the data holdings maintained by the STFC's laboratories (notably the Rutherford Appleton Laboratory and the Daresbury Laboratory) and also by various external projects in which the DMG participates.

Finally, Andrei Lopatenko has written a *Resource Guide to Metadata for Science and Research* [21], which gives information on many other additional resources, but unfortunately it appears not to have been updated since 2002. There are no specific standards for metadata to describe scientific software.

3.2.1 Dublin Core

The Dublin Core (or DC) is not in itself a scientific metadata standard. It is mentioned here because it forms the basis of several of the scientific metadata standards listed below. Indeed it is in widespread use and also forms the basis of metadata standards in various other disciplines. The standard emerged from librarianship, particularly work on interoperable on-

line library catalogues. It takes its name from Dublin, Ohio, where it was defined in 1995 at a meeting organised by the OCLC (Online Computer Library Center) [22]. The Dublin Core is now superintended by the DCMI (Dublin Core Metadata Initiative) [23] and is defined in standards ISO 15836 (2009) [24] and NISO Z39.85 (2007) [25]; in the context of digital curation see also Day (2005, p12).

The most basic level of the standard is the Simple Dublin Core, which is deliberately limited to just fifteen 'elements.' Each element is a name-value pair. The names of each element, and its meaning, are defined. Examples include 'title,' 'author,' and 'date.' Together the elements describe or annotate the dataset to which they refer. For a given dataset elements may be omitted (if they are not relevant) or repeated (if appropriate). The Qualified Dublin Core is a more sophisticated version of the standard which introduced three additional elements and other sophistications.

The Dublin Core standard defines the basic elements and their meanings. It deliberately does not define how they are represented or encoded. Various encodings are available and XML is now common.

3.2.2 DIF

Domain: scientific data sets.

DIF (Directory Interchange Format) [26] is a standardised format for exchanging information about scientific datasets. It is principally a discovery format: the information provided will allow users to determine whether a dataset is suitable for their purposes. DIF is one of the longest-established metadata formats, originating in the *Earth Science and Applications Data Systems Workshop* (ESADS) held in 1987, with the first version of the standard adopted in 1988.

DIF complements other metadata standards by specifying a 'container' to hold a set of metadata entries that describe a dataset. This set of entries comprises just eight mandatory entries and additional optional ones. Some of the entries can be free text, others have a controlled vocabulary.

DIF is not defined in terms of XML (indeed, it pre-dates it), but an XML schema is now available. DIF (including its schema) is maintained by the NASA Goddard Space Flight Center as part of the Global Change Master directory (GCMD) [27], a directory of Earth science data sets and related tools and services.

3.2.3 Darwin Core

Domain: biology.

The Darwin Core [28], often referred to as DwC, is a body of standards constituting a metadata specification for biological data. Specifically it is intended to facilitate the exchange of information about biological diversity. It allows the geographic occurrence of species and the existence of specimens or examples of these species in collections (either physical collections in museums or digital archives) to be represented. The system is principally based on taxa (that is, a group of related items in a classification scheme; recall the importance of classification in the biological sciences discussed in Section 2.1) and was originally intended to support the discovery, retrieval and integration of this information. The scope is now broader and the Darwin Core provides a standard mechanism for sharing information on biological diversity, with a particular emphasis on facilitating the reuse of information in different contexts.

The Darwin Core was originally based on the Dublin Core standard for library metadata (see Section 3.2.1, above) and the name is a deliberate acknowledgement of this provenance. The Darwin Core can be viewed as an extension of the Dublin Core for biodiversity. A glossary of standardised terms is an important part of the standard.

The full Darwin Core is extensive and flexible, but a more restricted, though common, way of using the format is the so-called 'Simple Darwin Core' [29]. The standard is also extensible through a name-space mechanism, enabling it to address additional biological disciplines. The standard has both XML and 'Fielded Text' (see Section 3.1.2 above) representations.

The Darwin Core is now an important international standard in widespread use. It is managed by Biodiversity Information Standards (TDWG) [30], which was previously the Taxonomic Database Working Group (and the earlier acronym is still retained). TDWG is a not-for-profit scientific and educational association that is affiliated with the International Union of Biological Sciences.

3.2.4 DDI

Domain: social science and archiving.

The DDI (Data Documentation Initiative) [31] is an effort to establish a standard for documentation describing social and behavioural science data in order to make such documentation interoperable. The DDI specification is written in XML and is used to create standard documents which can be presented in a variety of ways by appropriate processing of the DDI XML tags (this is exactly the use for which XML and its predecessor SGML were originally intended, of course).

Work on DDI started in 1995 and the latest version of the format, 3.1, was released in October 2009. The development of DDI is overseen by the Data Documentation Initiative Alliance [32] which has an international membership.

3.2.5 TEI

Domain: social sciences, linguistics and humanities.

The TEI (Text Encoding Initiative) [33] is a standard for representing texts, chiefly in the social sciences, linguistics and humanities. Work on the initiative started in 1987. There have been several versions of the standard since it was introduced in 1994. Early versions were expressed using SGML, version P4 had both SGML and XML representations and the current version, P5, has only an XML representation and is intended to take advantage of related standards such as XSLT and Xquery.

TEI defines a set of XML tags for marking up documents. The set is extensive, with about 500 elements, because TEI is intended to be able to represent any text from any period, though it is likely that only a much smaller set will be used in any given document. The TEI XML tags divide into broadly two types, one for representing details of the text (such as the author, bibliographic information, provenance *etc*; that is typical metadata details) and others for describing the structure of the document (sections, headings *etc*).

A large number of projects have used the standard and it has had an important impact on digital scholarship. Development is overseen by the international Text Encoding Initiative Consortium [34].

3.3 Standards for geospatial metadata

This section considers some of the more common standards for geospatial metadata. Geospatial datasets [ii] have the common property that they are related to a defined position with respect to the Earth's surface (often at that position, but also perhaps above or below it). This feature provides one obvious way of combining such datasets and has proved a key incentive to interoperability and metadata standardisation in this area. Several common metadata formats are listed in Table 2 and discussed briefly below. The DIF format mentioned in the previous section can also be used to represent geospatial data, albeit with some limitations. DwC, also mentioned above, is not a geospatial format but necessarily has some geospatial aspects. GML (Geography Markup Language) [35] should also be mentioned. It is

an XML-based language for describing geospatial data and so encompasses more than just geospatial metadata.

A number of distributed systems have been developed for linking dispersed Earth-science resources and archives. Typically such systems make extensive use of geospatial metadata to achieve interoperability. Two examples might be mentioned: GEON / GEONGRID [36] is developing a comprehensive geoinformatics system largely concerned with the solid Earth and the Earth System Grid [37] is mostly concerned with climate change simulations. Other examples include NEON [38]. LTER [39] and GEOSS [40].

3.3.1 ISO 19115

ISO 19115 is an international standard for geographic metadata developed by ISO technical committee TC 215 [41]. Its purpose is to provide a clear procedure for the description of digital geographic datasets. It achieves this aim by defining a common set of terminology, definitions and extension procedures for geographic metadata. It does not define a set of geospatial metadata items, nor an encoding scheme for such items. The first version of ISO 19115 was released in 2003. It attempted to harmonise earlier formal and *de facto* standards by defining a common terminology and set of definitions in which they could be expressed. New or existing formats have subsequently been respectively cast or re-cast as 'profiles' or recommended subsets of the standard. ISO 19115 is one of a family of geographic information standards overseen by ISO TC 215.

3.3.2 CSDGM

CSDGM (Content Standard for Digital Geospatial Metadata) [42] is a US standard for geospatial data. The standard was originally adopted in 1994 and the current version (FGDC 1998) is a revision dating from 1998. CSDGM provides a common terminology and set of definitions for the documentation of digital geospatial data in order to support the discovery, access and transfer of such data. There is a legal requirement on virtually all federal agencies in the US to use the standard to document the geospatial datasets that they hold. Though originally a US standard CSDGM is now in widespread use throughout the world. It is extensible: profiles can be defined which include additional elements suitable for a particular application area or type of dataset.

CSDGM is maintained by the US Federal Geographic Data Committee (FDGC) [43] and consequently is often referred to as the 'FDGC Metadata Standard'. The FDGC is leading the development of an ISO 19115 profile (see above) for CSDGM. In addition the NOAA Coastal Services Center has developed an XML representation for it.

3.3.3 INSPIRE

INSPIRE (Infrastructure for Spatial Information in the European Community; EC 2008) [44] aims to provide a spatial data infrastructure for the EU (European Union). It was introduced following EU Directive 2007/2/EC which came into force in May 2007 and must be fully implemented by 2019. INSPIRE aims to create an extensive spatial infrastructure and geospatial metadata underpins much of the initiative.

3.3.4 GEMINI

GEMINI [45] is the UK standard for discovery geospatial metadata. It defines a set of metadata elements for UK geospatial discovery-level metadata. (Recall that discovery-level metadata are metadata elements which allow datasets that are suitable for a given purpose to be identified.) The first version of the standard was released in 2004. The current version, GEMINI 2 (AGI 2009) was released in 2009. It has been revised to meet the requirements of the EU INSPIRE directive and to conform to the international standard ISO 19115 (see above for both). GEMINI 2 will form the basis of a UK geospatial metadata discovery service.

3.4 The fourth paradigm and the Semantic Web

Systems providing aspects of the 'fourth paradigm' of data-intensive science typically involve the automatic identification and subsequent searching of geographically-dispersed heterogeneous data archives and other resources. The results obtained from these various archives will be in diverse formats and must be automatically combined to form a unified report for the user. A full discussion of these systems is well beyond the scope of this chapter. However, typically they rely on metadata to facilitate inter-operation. The metadata in the various archives must be comparable, either by standardisation or, in the broadest sense, by the existence of rules for translating metadata items in one resource into those of another.

'Fourth paradigm' systems are an example of the 'Semantic Web' [46] and various systems have been developed under this rubric to mediate such metadata comparison. Two that are often encountered are RDF and OWL. RDF (Resource Description Framework) [47] is a family of specifications from the World Wide Web Consortium (W3C) that provides a data model for metadata. It is explicitly designed to facilitate data interchange on the Web. Though originally devised for modelling metadata it has also been used for more general modelling of information resources on the Web. OWL (Web Ontology Language) [48] is a family of languages for representing ontologies which is also endorsed by the W3C. In computing usage an ontology is a representation of all the entities in some domain of knowledge and the interrelations between them. OWL ontologies are characterised by formal semantics based on RDF and XML. The collaborative knowledge base Freebase [49] has some similarities with Semantic Web systems.

Despite the availability of systems such as RDF and OWL the automatic interoperability of disparate systems poses a formidable challenge. In particular, there may not be a perfect one-to-one mapping between separate schemas or ontologies covering similar or overlapping domains. There is often more than one way to classify or categorise the items in a domain and these various ways are not necessarily commensurable. The ISO 19115 profiles developed for CSDGM and GEMINI (above) are attempts to define mappings between geospatial metadata formats. Section 4.4, below, gives an example of this type of 'fourth paradigm' system.

4 Examples of Scientific Metadata

This section will introduce a few examples of the use of scientific metadata. The examples are deliberately diverse, particularly in terms of scale, and range from simple to more complicated systems.

4.1 Basic metadata: annotating an astronomical image

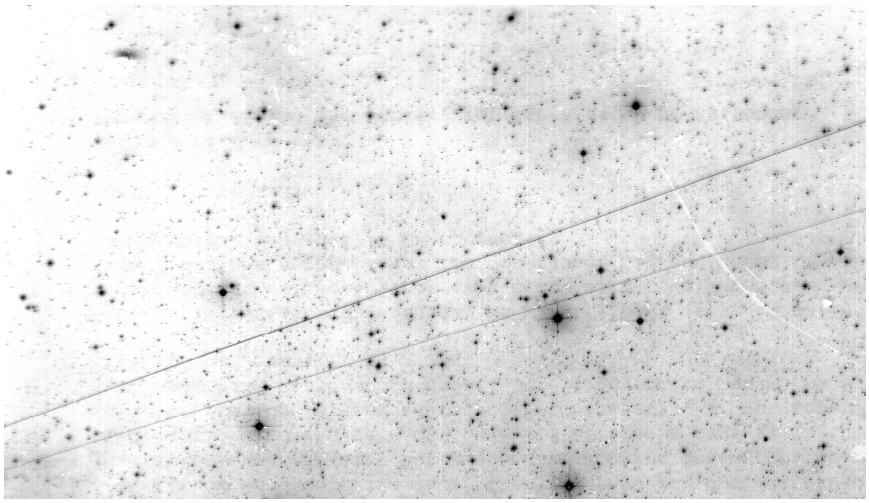


Figure 4: Photograph of a star field (courtesy the Royal Observatory Edinburgh).

As an example to illustrate the very basic use of metadata, Figure 4 shows a simple photograph of a patch of sky. Following the normal convention in astronomy it is shown as a negative image, so the stars appear dark against a bright sky. The photograph was taken by making a long exposure with a camera attached to a telescope. The telescope moved to compensate for the rotation of the Earth, so the stars remain as points rather than being stretched into arcs. The two prominent straight lines are trails due to artificial satellites that passed through the field of view. As it stands this image is completely useless. It is impossible to know even which part of the sky is being photographed. If the caption was reworded so that it read:

'Part of a photograph of Schmidt survey field 159 in the southern constellation of Pictor. It was taken by the UK Schmidt Telescope at the Anglo-Australian Observatory in New South Wales on 28 October 1976. The exposure time was 70 minutes.'

Then some basic details about the image are given, albeit in a form that can only be interpreted by a human. If the image was stored in the FITS format (see Section 4.2, below) then additional details describing it could be stored in the FITS headers (see Figure 5 (a)). A suitable programme reading the FITS file could interpret and use these details, but it must, of course, know the meaning of every header keyword.

FITS is a bespoke format. For wider interoperability the header keywords need to be expressed in a more standard fashion, typically XML, as shown in Figure 5 (b). XML and tools that operate on it are now ubiquitous, thus facilitating the widespread syntactic interoperability of this representation. However, for semantic interoperability, that is for diverse applications to understand the meaning of the values and use them correctly, then the XML tag names must have some agreed interpretation, perhaps through a shared schema or some sort of ontology. Ultimately such schemes require the specification of appropriate standards.

This example has discussed storing metadata in the headers of a FITS file. As an aside, some standard image formats, such as JPEG, also permit some header (or metadata) items to be included in an image and modern digital cameras will often use this facility to automatically include, for example, the time when an image was taken.

4.2 The FITS data format

FITS (Flexible Image Transport System) is the most widely used data format in astronomy. It is described here as an example of a data format which allows metadata, in the form of header information, to be stored with the data that they describe. FITS is principally designed to store n-dimensional arrays but it can also store other types of data such as tables. Though it is not specifically an image format it can store images (which are just two-dimensional arrays). Consequently it is sometimes encountered outside astronomy and some conventional image display applications can read it. The FITS format was originally defined in 1981 and a number of extensions have subsequently been added. Deliberate decisions were taken early in the history of the standard that (i) it should define the format down to the level of individual bits and (ii) that the standard, and its extensions, should be published in the primary, refereed astronomical literature. These decisions were taken to ensure, as far as practical, that data written using the format would remain readable. The original standard is defined by Wells *et al*. (1981). It is now maintained by the FITS Support Office at the NASA Goddard Space Flight Center [50].

The full details of the FITS format are not germane here and only the way that metadata can be accommodated will be outlined. There can be several datasets in a file and each dataset is preceded by a header. This header comprises one or more 2880-byte blocks consisting of ASCII characters. Each block is divided into 36 header records, each of 80 characters [xiv].

Every header record contains a keyword for which the basic form is:

name = value / optional comment

(see Figure 5(a) for examples). There are restrictions on the length of the name and the characters that it may contain and on the position of the other items. However, the basic principle is that a set of named keywords specify the details of the dataset.

```
PLATENO =J2673
FIELD   =159
RA_CEN  =05:30.0
DEC_CEN =-55:00
EQUINOX =J2000
EPOCH   =1976-10-28
TELESCOP=UKST
EXPOSURE=70
```

(a)

```
<PLATE>
  <PLATENO>J2673</PLATENO>
  <FIELD>159<FIELD>
  <PLATE_CENTRE>
    <RA>05:30.0</RA>
    <DEC>-55:00</DEC>
    <EQUINOX>J2000</EQUINOX>
  </PLATE_CENTRE>
  <EPOCH>1976-10-28</EPOCH>
  <TELESCOPE>UKST</TELESCOPE>
  <EXPOSURE>70</EXPOSURE>
</PLATE>
```

(b)

Figure 5: Metadata for the image shown in Figure 4: (a) as FITS headers and (b) in XML.

Keywords may be mandatory, reserved or optional. Mandatory keywords (for example, SIMPLE, BITPIX and NAXIS) must be present. They specify the basic properties of the dataset, are defined as part of the standard and must be used as described in the standard. Reserved keywords are also part of the standard. They are optional but if present must be used as described in the standard. Optional keywords may be freely invented, provided that their names conform to the naming rules. An arbitrary number of optional keywords may be added to a dataset to specify any required metadata. Programmes reading the FITS file may access these metadata to interpret the dataset. However (and crucially) the meaning assigned to these items is not part of the FITS standard and must be agreed (with a lesser or greater degree of formality, as appropriate for the circumstances) amongst the groups and institutions involved.

There are keywords for specifying the units of the dataset (for example, BUNIT for arrays) and recommendations about the units preferred for various quantities, adapted from the recommendations of the International Astronomical Union and ultimately largely based on the SI system. Keywords themselves, however, are defined by three attributes: a name, a value and optionally a comment. There is no standard way of specifying the units of a keyword. Thus, any units for a keyword must be explicitly stated as part of its agreed definition. So, for example, keyword RADVEL might be defined as the 'heliocentric radial velocity in km/sec with recessional velocities positive' rather than ' heliocentric radial velocity'. This circumstance does not matter for most of the mandatory and reserved keywords, which are dimensionless (that is, do not have units). Some alternative data formats allow a richer set of attributes for keywords performing a similar function, which may include the specification of units. For example, Table 3 shows the keyword attributes permitted for the star catalogue manipulation system CURSA (Davenhall 2001).

In addition to named keywords FITS headers can also contain COMMENT and HISTORY header records. An arbitrary number of both types of record can be included. Comments are intended to contain descriptive text that can be read by a human. HISTORY records were an early attempt to include provenance information describing the origin and processing history of the dataset. However, the HISTORY record is defined to contain free-format text and, like the COMMENTS records, is intended to be read by humans. Standardised, automated details

can be included, of course, but any such schemes must be locally agreed and will only be understood amongst the participating groups.

4.3 NanoCMOS circuit simulation

The EPSRC pilot project *NanoCMOS: Meeting the Design Challenges of nano-CMOS Electronics* (Sinnott *et al*., 2006; Sinnott *et al*., 2007; Reid *et al*., 2009) is investigating the challenges for CMOS (Complementary Metal–Oxide Semiconductor) microprocessor design posed by the decreasing size of individual transistors. It is outlined here as an example of a project which is using a catalogue of metadata to summarise a large body of results.

Briefly, the background to this work is that it is a commonplace that smaller and more capable CMOS devices become available as time progresses (and consequently computers become more powerful). This circumstance is captured in 'Moore's Law' which has applied loosely since the integrated circuit was invented in 1958. The 'happy scaling' encapsulated in Moore's Law has led to the enormous success of the computer industry, but it may be coming to an end. CMOS devices consisting of individual transistors 40 nm across are already in mass-production. Transistors smaller than 10 nm are anticipated by 2018. A single silicon atom is approximately 0.1 nm in diameter and thus 10 nm is equivalent to a line of about a hundred silicon atoms. For comparison the wavelength of visible light is about 500 nm. Unsurprisingly, there are major challenges in further reducing the transistor size.

Historically circuit and system design has treated the individual transistors comprising a device as being uniform and similar. As the size of the transistors shrinks this assumption no longer holds; the variability between transistors increases with decreasing size. This effect has several causes, some fundamental to the quantisation of charge and matter. Other important effects include the disposition of dopants in the semi-conductor and the roughness of the edge of the transistor. These effects are stochastic and cannot be mediated by better manufacturing process control.

Though this variability between transistors is inherent it can be characterised. Circuits and systems can then be designed in a way that accommodates it. To perform this characterisation the NanoCMOS engineers use a layered suite of application programmes (Figure 6). The arrow in the diagram represents increasing device complexity. At the bottom are applications to model individual transistors and characterise their range of properties. Applications in the higher layers are concerned with designing individual circuits and then agglomerations of circuits into systems. Typically each application will use data from lower applications. Figure 6 encompasses a considerable range of complexity; large, modern VLSI (Very Large Scale Integration) systems may contain several thousand million transistors.

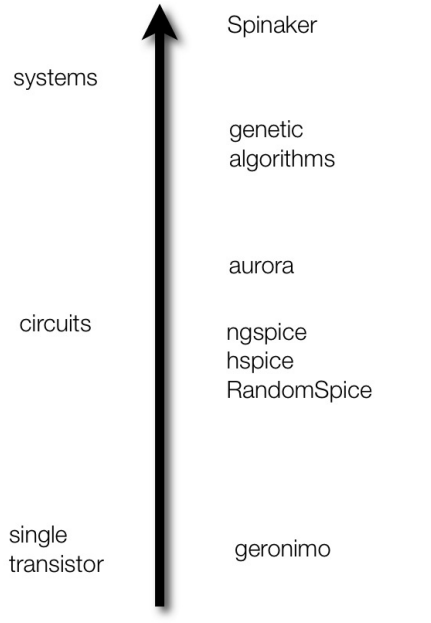


Figure 6: The increase in complexity in CMOS devices from single transistors to circuits and then systems. The middle column lists relevant simulation programmes and systems. The NanoCMOS group working in each area is shown on the right

The variability of individual transistors, and the consequent effect on circuits and systems, is addressed by performing simulations. Simulations are carried out on a variety of levels, from individual transistors to systems. Moreover, even simulating a single transistor is a complex operation requiring considerable processing time. Further, each simulation must be run many times for a range of input values in order to characterise the variability. Simulations of circuits will use the results generated for individual transistors, and simulations of systems will, in turn, use the results for circuits.

Assembling all the required simulations involves running many thousands of jobs, each typically with many sub-jobs, and generating, in total, multiple Tbytes of simulated data. There are far too many jobs to keep track of manually. Instead, in order to manage the jobs and make the results of the simulations available as required NanoCMOS has implemented a Data Management System (DMS) to capture the metadata for each job, store it and make it available as required. Typically the metadata for each job comprises the details specified to run the job (mostly the parameters for the simulation), a summary of the results and pointers to the actual results files.

The data files are stored using the OpenAFS [51] implementation of the AFS distributed file system (Howard *et al*, 1988) which allows convenient geographically dispersed access while providing fine-grained control over access rights. The metadata for each job are imported into the DMS while the job runs and stored separately in a DBMS. The metadata is expressed using XML.

The NanoCMOS project is still in progress. The metadata stored for each type of simulation are defined by the scientists and engineers performing the simulations, not the computer specialists implementing the DMS. This approach is obviously correct: only the domain specialists have the necessary knowledge. Further, the approach is pragmatic and incremental; the NanoCMOS techniques are new and the domain specialists simply do not know all the details that are required, so additions and changes are made as they are found to be necessary. Similarly, because the system is new the various metadata items are not completely documented. However, this situation will improve as the project moves from development to production, the metadata becomes more stable and a body of simulations are generated that need curation. It does, however, illustrate the common problem of identifying all the metadata necessary to adequately annotate scientific datasets.

4.4 The Virtual Observatory

The astronomer's 'virtual observatory' is an ambitious idea that has been developed since the beginning of the present century. It is described here as an example of the realisation of the 'fourth paradigm' of data-driven science for a single discipline. The basic aim of the Virtual Observatory is to provide seamless, simultaneous access to a geographically dispersed collection of astronomical data archives. It was initially driven by the development of several new sky surveys that produced datasets of hitherto unprecedented size. However, the aim also encompassed access to smaller and more specialised data archives, the journal literature and direct access to, and automatic scheduling of observations on, some telescopes (to detect and secure additional observations of transient phenomena).

Various projects have been funded in several countries to contribute to aspects of the virtual observatory, and these co-operate under the umbrella of the International Virtual Observatory Alliance (IVOA) [52]. A great deal of work has been done on specifying the architecture and implementing applications. The components are necessarily distributed and communicate by Web services and XML.

However, more pertinent to the present discussion, the IVOA has developed standards to specify the interaction between components. In particular, data archives and similar services which wish to participate in the VO must ensure that their datasets are annotated with suitable

metadata. In order to permit inter-operation the metadata must have semantic as well as syntactic standardisation: the items must have a common meaning as well as a common syntax. The difficulty of establishing such a set of comprehensive definitions, even within a single discipline, should not be underestimated. Achieving consensus typically requires much careful discussion of the technicalities. Though progress has been made, it has been slow and the work continues.

Similar systems are being developed in a variety of disciplines. One important one is GEOSS (Global Earth Observation System of Systems) [40].

5 The Digital Curation of Scientific Metadata

This section considers the digital curation of scientific metadata. Metadata, scientific or otherwise, are not, of course, curated in isolation. Rather, they will be curated along with the data that they are associated with and in some cases the processes that create them. Or more precisely, it is the data that are being curated and the scientific (and other) metadata are included to allow proper access to and interpretation of these data. Usually the curation and preservation will be occurring in the context of retaining the data in an archive (as discussed in Section 2.2 above): most likely a data archive, but perhaps a journal archive or a project or institutional archive.

The details will vary enormously in terms of scope, scale and purpose between different projects and different archives, and this section can only make the most general comments. Further, normal curation practices and procedures will apply to the archive, as described in, for example, the DCC's own *Digital Curation Lifecycle Model* (DCC 2008; Constantopoulos *et al*. 2009) or perhaps the OAIS preservation reference model (Day 2005, p19, CCSDS 2002). This section will not duplicate the advice in these documents. However, it is worth briefly recalling a few pertinent questions about the construction of a data archive:

- Why are the data being curated and who will access them?
- For how long are the data required?
- What access to the data is required?
- What strategy is anticipated to facilitate the required access for the required length of time?

The nub of the problem of curating scientific metadata is to ensure that all the metadata (and its associated documentation) that are required to use the data are captured in the archive and are accessible to their legitimate users. Who will use the data and for how long are critically important here. Recall that in many scientific projects the original user communities are small and specialised. If the users of the archive are this same community and their close colleagues, and retention is only anticipated in the short term, then there may be a temptation to capture less metadata and write only brief documentation. This approach should only be adopted, if at all, with very considerable caution: it is relying on researchers being able to remember details and being available to communicate with colleagues. In practice memories are fallible and short-term funding mechanisms often result in a high and rapid turn-over of staff. Conversely, if the users are a wider community of non-experts or use is required over a longer period of time then much greater care needs to be taken to ensure that every required item of information is captured and documented.

In this latter case it is necessary to ensure not just that all the required metadata values are captured, but that documentation (in the broadest sense) that allows the metadata items to be understood is also captured. Quantities must be defined and, if appropriate, their units specified. It may be necessary to collect and curate documents from the local 'grey literature' and to preserve references (or links) to refereed journal articles describing the experiments or datasets.

In some cases it may be necessary to curate software to access the data, together with instructions on how to use this software and notes on the environment in which it should be run. Most scientific data are generated, and archives operated, under tight budgetary and time constraints. It will rarely be feasible for the archive to provide a full emulation of the software's original environment. Conversely, much scientific data processing software will run under a fairly basic Unix operating system.

It will be apparent that these tasks require a deep understanding of the data being curated, the software to process and access them and the scientific context in which the work is being done. Also this discussion is veering into the curation of scientific data rather than scientific metadata. But this digression is inevitable: the only point of curating the metadata is to make the data being curated accessible and interpretable.

5.1 Eight questions

This section lists eight questions that it might be useful to answer when considering the curation of scientific metadata.

1) Do you have a deep understanding of the application domain and the data being curated and / or access to colleagues with such expertise, or ideally both?

2) Are all the metadata necessary to process and access the data included? This question is a catch-all which is easier to ask than answer. One simple test might be to ensure that users unconnected with the construction and operation of the archive can identify data of interest in it, extract them and then process them. Beware that in scientific archives processing the data to do useful work will probably involve doing more than simply displaying the data.

3) Are the metadata (and data) adequately documented? Recall that 'hidden assumptions' and 'common or shared knowledge and vocabulary' are common amongst the original investigators who generated the data. If necessary persuade them to write any documentation which is missing. Ensure that any links or references to the refereed literature or external documentation are recorded. If there is any doubt about the permanence of external documentation then obtain, keep and curate local copies and ensure that they are accessible.

4) What format are the metadata and data in? Is the format a standard? If so, who maintains the standard and (if appropriate) which version is in use? If the metadata (and data) are held in a local format is the format itself (as distinct from the data and metadata it holds) documented?

5) What software is needed to access the data and metadata? Does any local software need to be curated? If so, is documentation available for this software?

6) Are the metadata kept separately from the data? If so, are adequate arrangements in place to link the various datasets with their associated metadata? If the metadata are stored in a DBMS (*cf.* NanoCMOS; see Section 4.3) how are the contents of the working DBMS to be archived and subsequently used?

7) Are the data in the archive to inter-operate with external archives or systems? If so are the metadata (and software) required to support this inter-operation in place?

8) Do you have a deep understanding of the application domain and the data being curated and / or access to colleagues with such expertise, or ideally both? This question, of course, is just a repeat of the first, but the point bears reiterating.

Most of these questions need to be addressed early in the curation process, when the data archive is being designed and the data input procedures developed. Some, however, continue throughout the lifetime of the archive. For example, if external metadata formats are being used these may evolve and need to be tracked. Note also that software to process the data may continue to develop after the generation of data has ceased. Finally, the effort required to keep the archive current should not be underestimated.

6 Discussion

This chapter of the DCC *Digital Curation Manual* has discussed scientific metadata. Scientific metadata typically provides additional information necessary to understand, analyse and interpret scientific data sets. There is, however, no strict differentiation between scientific and 'non-scientific' metadata. A scientific data set being curated will typically require standard metadata, for example preservation metadata, as well as more obviously scientific items. There is also no clear distinction between scientific data and scientific metadata. Many scientific data formats, for example, have provision to include metadata items.

Scientific (and other) metadata are curated together with the data they annotate. It is meaningless to discuss their curation in isolation from that of the data they describe. Some basic points to bear in mind about scientific metadata and their curation are:

- Scientific data are generated by experiments and observations as part of the scientific process. That is, they are ultimately experimental, rather than routine. This point is less important for large, long-lived, collaborative experiments, but is ultimately inherent to the scientific process.

- Scientific metadata are likely to be more extensive and less standardised than non-scientific metadata.

- Scientific datasets are often generated with incomplete metadata. There can be 'hidden knowledge and shared assumptions' amongst the experimenters generating the data. Calibration values and other information may be shared informally 'on the back of an envelope.' Similarly the metadata (and data) items in the dataset are not always adequately documented or their units, if any, specified. A challenge of curating such data is to ensure that all the necessary auxiliary information and documentation is collected and retained, along with the data themselves.

- Scientific user-communities are often small and specialised. If data are to be used outside their original communities, or preserved for an extended period of time, additional metadata and documentation may be required.

- Standards, both syntactic and semantic, are needed to facilitate interoperability and re-usability. Physical quantities need precise, and documented, definitions and numeric values must have known units.

- Standards may be specific to the specialised community that generated the data. Because the communities are small standards (and other practices) can evolve rapidly and so must be tracked.

Scientific data are often held in specialised formats such as HDF, NetCDF, DICOM *etc*. (see Section 2.5). Many of these formats have facilities for including metadata. Thus, a dataset and its associated metadata will reside in a single file. This arrangement has the considerable advantage that the metadata and data are kept together and hence the metadata are available when the data are accessed or processed. However it is less suitable where an archive is being searched to identify datasets which match a query. Each file must be accessed to check its metadata items. To avoid this problem data archives will often keep a copy of the metadata (or

a subset of them) in a database or metadata registry which can be searched conveniently and expeditiously. Obviously if two such copies of the metadata are retained then they must be kept in step.

Often researchers accessing scientific data archives need to make specialised searches. Searches of geospatial datasets based on geographical location are one common example. However, it is usual for researchers to need to make bespoke, one-off queries on some combination of a wide variety of metadata items. Consequently scientific data archives typically require powerful and flexible query facilities. Similarly, the 'fourth paradigm' of data-intensive science depends on datasets being annotated with syntactically and semantically correct metadata which can be automatically searched to permit inter-operation.

Despite these various challenges, retaining adequate metadata is central to the curation of scientific data: it is necessary if the data are to remain useful. In *Two Choruses from 'the Rock'* T.S. Eliot asks:

> *Where is the wisdom we have lost in knowledge?*
> *Where is the knowledge we have lost in information?*

In the sciences metadata can allow information, or data, to be turned into knowledge. Wisdom, alas, remains elusive.

Acknowledgements

I am grateful to Yin Chen, Mike Mineter and two anonymous referees for helpful comments on an earlier version of the manuscript. Any mistakes, of course, remain my own.

Notes

[i] The *Oxford English Dictionary* is instructive about the origin of the term 'metadata'. The first use that it records is as 'meta data' (two words) in the *Proceedings of the IFIP Congress 1968* which was held in Edinburgh and published in 1969.

The prefix 'meta' is of Greek origin and has its usual meaning (in the sciences) of: 'above, beyond, at a higher level or encompassing'. 'Data' is, of course, of Latin origin, the plural of 'datum', which in English usage originally meant a known or assumed fact that was given or granted. It is in this sense that in *A Scandal in Bohemia* Sherlock Holmes reminds Dr Watson that 'it is a capital mistake to theorise before one has data'. The modern usage of 'a body of usually numeric values' is surprisingly recent, dating only from the end of the nineteenth century.

'Metadata', then, is a portmanteau word, half Latin and half Greek, and, as was once said of 'television' for the same reason, 'no good can come of it.'

[ii] Geospatial data are geographic or other environmental data which are referenced by geographical location. The location will be defined by coordinates which ultimately correspond to terrestrial latitude and longitude, though they may be expressed in some other system, such as the Ordnance Survey Grid Reference used in the British Isles.

A geospatial system is a system designed to process geospatial data.

[iii] Quantum Mechanics, and in particular the Uncertainty Principle, is a quantification of the extent to which the world is predictable and repeatable: the 'rigidly defined areas of doubt and uncertainty' once famously demanded by the Amalgamated Union of Philosophers, Sages, Luminaries and Other Thinking Persons.

[iv] For example, the standard MK spectral classification for stellar spectra is a 'good' scheme because the classification categories correspond to a physical parameter: the surface temperature of the star. Conversely, the standard 'Hubble tuning fork' classification scheme for galaxies is less useful because the classification categories do not obviously correspond to any single physical characteristic of the galaxy.

[v] Indeed one of the characteristics which distinguishes science from its antecedents such as alchemy is the importance placed on openly reporting and discussing results and the general emphasis on communication amongst participants.

[vi] This chapter is not the place to discuss the history of scientific societies. However, briefly, the earliest association of individuals interested in what would now be called the sciences appears to have been the Polish *Sodalitas Litterarum Vistulana*, which was founded in Cracow as early as 1488. However, scientific societies along recognisably modern lines are contemporary with the Scientific Revolution of the seventeenth century and include the Italian *Accademia dei Lincei* (1603), the *Académie Française* (1635), the *Deutsche Akademie der Naturforscher Leopoldina* and the Royal Society of London (1660). The earliest scientific journal was the French *Journal des sçavans* which began publication in January 1665, though it later foundered. The Royal Society's *Philosophical Transactions* started a couple of months later and publication has continued uninterrupted to the present.

[vii] Preprints are copies of papers formally published in journals which are circulated privately amongst groups of collaborating institutions or individuals. The system is of long-standing and formerly paper copies were circulated by mail. Copies are now largely circulated electronically and may be made publicly available rather than circulated to a closed list. Indeed, many disciplines maintain publicly accessible archives of preprints. The importance of preprints is that they allow smaller institutions, with limited library budgets, and their staff, to keep informed about recent research.

[viii] There is now a formally-published, peer-reviewed academic journal devoted to the study of grey literature: *The Grey Journal: An International Journal on Grey Literature* [53]. It seems a particular irony that there should be a formal publication devoted to studies of the unofficial *samizdat* grey literature.

[ix] A relatively small number of distributed scientific applications are run by interested members of the public on their home computers. SETI@Home is an early and well-known example. To an extent these applications are a counter-example to the small user base for scientific software, as they are run by large numbers of people. However, there are still relatively few such applications and, moreover, the number of users is still much less than for common PC office or home software.

[x] Some of the *voyages extraordinaires* stories of the French author Jules Verne (1828-1905) provide an unusual example, particularly *From the Earth to the Moon* (1865) and *Around the Moon* (1873). Verne has long had a mixed reputation in English-speaking countries, in part because he was ill-served by his early translators. They were convinced that he was an author of children's stories and translated accordingly. However, more pertinently, they were not familiar with both the metric system and Imperial units. Where units were specified they would often simply replace Verne's metric units with the corresponding Imperial ones, but without changing the numeric value, so, for example, 100 km/hour would become 100 mph, thus rendering Verne's careful calculations into gibberish.

[xi] MARC (MAchine Readable Cataloguing) [54] is a library standard for representing bibliographic and related information. It was originally developed in the 1960s and is now well-established. It is widely used by libraries and forms the basis of many library catalogues. The record structure used by MARC is an implementation of the library standard ISO 2709 (ISO 2008), which is also known as ANSI/NISO Z39.2.

[xii] The MIME (Multipurpose Internet Mail Extensions) standard was developed to extend the email format to include features such as character sets other than ASCII, attachments in various formats and header information which is not ASCII. However, the format is also used in the Web HTTP protocols (where it is usually referred to as the MIME-type). MIME allows additional header fields to be defined for email messages *etc*. Each field is defined as a name (or 'key') / value pair. MIME is defined in a set of six RFCs: 2045-49. The details are not germane here, but a couple of overviews are available [55, 56].

[xiii] DIME (Direct Internet Message Encapsulation) was a standard for streaming binary and other encapsulated data over the Internet that was proposed by Microsoft in the early 2000s. It saw some limited use, but in 2002 work on the standard was suspended and the draft RFC which would have specified it was withdrawn. A couple of overviews are still available [57, 58].

[xiv] Thus each header record is a 'card image', that is it contains the same number of characters as a standard punched card. This property was considered important when the format was defined.

References

All URLs checked on 23 January 2011.

AGI, 2009, *UK GEMINI Standard*, version 2, Association for Geographic Information. See: http://location.defra.gov.uk/wp-content/uploads/2009/12/GEMINI2.pdf

Atkinson, M., 2009, *Research Data: It's What You Do With Them* (guest editorial), *The International Journal of Digital Curation*, **4**(1).

van Ballegooie, M., and Duff, W., 2006, *Archival Metadata*, chapter in *DCC Digital Curation Manual*, Digital Curation Centre (DCC), Edinburgh.

BIPM, 2006, *Le Système international d'unités / The International System of Units* (eighth edition), Paris: BIPM (Bureau international des poids et mesures)

Campbell, L, 2007, *Learning Object Metadata*, chapter in *DCC Digital Curation Manual*, Digital Curation Centre (DCC), Edinburgh.

Caplan, P., 2006, *Preservation Metadata*, chapter in *DCC Digital Curation Manual*, Digital Curation Centre (DCC), Edinburgh.

CCSDS, 2002, *Reference model for an Open Archival Information System (OAIS)*, CCSDS 650.0-B-1, Consultative Committee on Space Data Systems, Washington, DC. See http://public.ccsds.org/publications/archive/650x0b1.pdf

Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y. and Papatheodorou C., 2009, DCC&U: An Extended Digital Curation Lifecycle Model, *DCC&U: An Extended Digital Curation Lifecycle Model*, *The International Journal of Digital Curation*, **4**(1).

Davenhall, A.C., Kelly, B.D., and Beard, S.M., 1984, *A User's Guide to the Haggis System for Handling Parametrised Data*, Royal Observatory Edinburgh.

Davenhall, A.C., *Database Applications in Starlink*, in *Databases and On-line Data in Astronomy*, Albrecht, M.A. and Egret, D. (eds), Kluwer, Dordrecht, pp165-78.

Davenhall, A.C., 2001, *CURSA – Catalogue and Table Manipulation Applications*, SUN/190.10, Starlink, Rutherford Appleton Laboratory, Oxfordshire.

Davenhall, A.C., Clowes, R.G. and Howell, S.B., 2001, *CURSA – A Package for Manipulating Astronomical Catalogues*, in *The New Era of Wide Field Astronomy*, Clowes, R.G., Adamson, A. and Bromage, G. (eds), Astron. Soc. Pacific, San Francisco (Astron. Soc. Pacific Conference Series) **232**, pp314-316.

Day, M., 2005, *Metadata*, chapter in *DCC Digital Curation Manual*, Digital Curation Centre (DCC), Edinburgh.

DCC, 2008, *The DCC Curation Lifecycle Model*, Digital Curation Centre. See http://www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf

Dirks, L., 2009, *Introduction* (to Part 4: Scholarly Communication) of *Hey et al*., 2009, *op. cit*., pp175-6.

European Commission, 2008, *Commission Regulation (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance)*. See http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32008R1205:EN:NOT

FGDC (Federal Geographic Data Committee, Metadata Ad Hoc Working Group), 1998, *Content Standard for Digital Geospatial Metadata* (CSDGM), FGDC-STD-001-1998, Federal Geographic Data Committee Secretariat, c/o U.S. Geological Survey, Reston, Virginia. See http://www.fgdc.gov/standards/projects/FGDC-standards-projects/metadata/base-metadata/v2_0698.pdf

Greenberg, J., White, H.C., Carrier, S. and Scherle, R., 2009, *A Metadata Best Practice for a Scientific Data Repository*, *Journal of Library Metadata*, **9**, p194.

Goble, C. and De Roure, D., 2008, *Curating Scientific Web Services and Workflow* , *EDUCAUSE Review*, **43**(5). See http://www.educause.edu/EDUCAUSE%2BReview/EDUCAUSEReviewMagazineVolume43/CuratingScientificWebServicesa/163168

Golinski, J., 2007, *British Weather and the Climate of Enlightenment*, Chicago Univ. Press.

Henry, J., 2002, *The Scientific Revolution and the Origins of Modern Science* (second edition), Basingstoke, Hampshire: Palgrave (Studies in European History series).

Hey, T., Tansley, S. and Tolle, K. (eds), 2009, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research. See http://research.microsoft.com/en-us/collaboration/fourthparadigm/

Howard, J.H., Kazar, M.L., Nichols, S.G., Nichols, D.A., Satyanarayanan, M., Sidebotham, R.N., & West, M.J., 1988 (February), *Scale and Performance in a Distributed File System*, *ACM Transactions on Computer Systems*, New York: Association for Computing Machinery, **6** (1), pp51-81.

ISO, 2004, ISO 8601: *Data Elements and Interchange Formats - Information Interchange - Representation of Dates and Time*, Geneva: International Organization for Standardization.

ISO, 2008, ISO 2709: *Format for information exchange*, Geneva: International Organization for Standardization.

Jaschek, C., 1989, *Data in Astronomy*, Cambridge Univ. Press.

Klyne, G. and Newman, C., 2002, *Date and Time on the Internet: Timestamps*, RFC 3339, Internet Engineering Task Force, Fremont, California. See http://tools.ietf.org/html/rfc3339

Love, S., 2009, *No One Peer-Reviews Scientific Software*. See http://chicagoboyz.net/archives/10436.html

Lynch, C., 2009, *Jim Gray's Fourth Paradigm and the Construction of the Scientific Record*, in *Hey et al*., 2009, *op. cit*., pp177-183.

Nuin, P., 2008, *How to improve scientific software?* See http://blindscientist.genedrift.org/2008/04/23/how-to-improve-scientific-software/ ('Blind Scientist'),

Okasha, S., 2002, *Philosophy of Science*, Oxford Univ. Press (Very Short Introduction series)

Reid, D., Sinnott, R.O., Millar, C., Roy, G., Roy, S., Stewart, Gordon, Stewart, Graeme & Asenov, A., 2009 (July), *Enabling Cutting-edge Semiconductor Simulation through Grid Technology*, *Journal of the Philosophical Transactions of the Royal Society A*, London, **367**, pp2573-2584.

Shiflet, A.B and Shiflet, G.W., 2006, *Introduction to Computational Science: Modeling and Simulation for the Sciences*, Princeton Univ. Press.

Sinnott, R.O., Asenov, A., Berry, D., Furber, S., Millar, C., Murray, A., Pickles, S. , Roy, S., Tyrell A. & Zwolinski. M., 2006 (September), *Meeting the Design Challenges of nanoCMOS Electronics: An Introduction to an EPSRC Pilot Project*, UK e-Science All Hands Meeting, Nottingham UK.

Sinnott, R.O., Asenov, A., Brown, A., Millar, C., Roy, G., Roy, S. Stewart, G., 2007 (September), *Grid Infrastructures for the Electronics Domain: Requirements and Early Prototypes from an EPSRC Pilot Project*, UK e-Science All Hands Meeting, Nottingham, UK (best paper award).

Sinnott, R.W. (ed), 1988, *NGC 2000.0*, Cambridge Univ. Press.

Sufi, S. and Matthews, B., 2004, *CCLRC Scientific Metadata Model: Version 2*, DL Technical Reports DL-TR-2004-001, Daresbury Laboratory, Warrington. See http://epubs.cclrc.ac.uk/work-details?w=30324

Warren-Smith, R.F. and Lawden, M., 1999, *HDS: Hierarchical Data System, Version 4.3, Programmer's Manual* (SUN/92.12), Starlink Project, Rutherford Appleton Laboratory, Didcot, Oxfordshire. See http://starlink.jach.hawaii.edu/docs/sun92.htx/sun92.html

Wells, D.C., Greisen, E.W. and Harten, R.H., 1981, *Astron. Astrophys Suppl*, **44**, p363.

Winsberg, E, 2010, *Science in the Age of Computer Simulation*, Chicago Univ. Press.

URLs

All URLs checked on 23 January 2011.

[1] (METS): http://www.loc.gov/standards/mets/
[2] (Swan Upping): http://www.royal.gov.uk/RoyalEventsandCeremonies/SwanUpping/SwanUpping.aspx
[3] (*Journal of Biological Databases and Curation*): http://database.oxfordjournals.org/
[4] (RCSB Protein Data Bank): http://www.rcsb.org/pdb
[5 (Scientific data format FAQ): http://www.cv.nrao.edu/fits/traffic/scidataformats/faq.html
[6] (File formats): http://en.wikipedia.org/wiki/List_of_file_formats
[7] (HDF): http://www.hdfgroup.org/
[8] (HDF-EOS): http://hdfeos.net/
[9] (NetCDF): http://www.unidata.ucar.edu/software/netcdf/
[10] (DICOM): http://medical.nema.org/
[11] (NEMA): http://www.nema.org/
[12I] (Medical Image Format FAQ): http://www.dclunie.com/medical-image-faq/html/
[13] (BIPM): http://www.bipm.org/en/si/
[14] (ISO 8601): http://www.iso.org/iso/support/faqs/faqs_widely_used_standards/widely_used_standards_other/date_and_time_format.htm
[15] (IETF; Internet Engineering Task Force) http://www.ietf.org/
[16] (ISO/IEC 11179): http://metadata-standards.org/11179/
[17] (W3C): http://www.w3.org/
[18] (XML): http://www.w3.org/XML/
[19] (Fielded Text): http://www.fieldedtext.org/
[20] (STFC Data Management Group): http://www.e-science.stfc.ac.uk/organisation/scientificapplications/data-management/datamanagement.html
[21] (*Metadata Resource Guide*): http://derpi.tuwien.ac.at/~andrei/Metadata_Science.htm
[22] (OCLC): http://www.oclc.org/uk/en/global/default.htm
[23] (DCMI): http://dublincore.org/
[24] (ISO 15836): http://www.iso.org/iso/iso_catalogue/catalogue_ics/catalogue_detail_ics.htm?csnumber=52142
[25] (NISO Z39.85): http://www.niso.org/kst/reports/standards?step=2&gid=&project_key=9b7bffcd2daeca6198b4ee5a848f9beec2f600e5
[26] (DIF): http://gcmd.gsfc.nasa.gov/User/difguide/difman.html
[27] (GCMD; Global Change Master Directory) http://gcmd.nasa.gov/
[28] (Darwin Core): http://rs.tdwg.org/dwc/index.htm
[29] (Simple Darwin Core): http://rs.tdwg.org/dwc/terms/simple/index.htm
[30] (TDWG): http://www.tdwg.org/
[31] (DDI): http://www.ddialliance.org/
[32] (DDIA): http://www.ddialliance.org/alliance
[33] (TEI): http://www.tei-c.org/index.xml
[34] (TEIC): http://www.tei-c.org/About/
[35] (GML): http://www.opengeospatial.org/standards/gml
[36] (GEON): http://www.geongrid.org/index.php
[37] (ESG): http://www.earthsystemgrid.org/
[38] (NEON): http://www.neoninc.org/
[39] (LTER): http://www.lternet.edu/
[40] (GEOSS): http://www.earthobservations.org/
[41] (ISO TC 215): http://www.isotc211.org/
[42] (CSDGM): http://www.fgdc.gov/metadata/csdgm/
[43] (FDGC and metadata): http://www.fgdc.gov/metadata/geospatial-metadata-standards
[44] (INSPIRE): http://inspire.jrc.ec.europa.eu/
[45] (GEMINI): http://www.gigateway.org.uk/metadata/standards.html
[46] (Semantic Web): http://semanticweb.org/wiki/Main_Page

[47] (RDF): http://www.w3.org/RDF/
[48] (OWL): http://www.w3.org/TR/owl-features/
[49] (Freebase): http://www.freebase.com/
[50] (FITS Support Office): http://fits.gsfc.nasa.gov/
[51] (OpenAFS): http://www.openafs.org/
[52] (IVOA): http://www.ivoa.net/
[53] (The Grey Journal): http://www.greynet.org/thegreyjournal.html
[54] (MARC): http://www.loc.gov/marc/
[55] (MIME): http://www.mhonarc.org/~ehood/MIME/
[56] (MIME): http://www.helpdesk.umd.edu/topics/email/protocols/315/
[57] (DIME): http://msdn.microsoft.com/en-us/magazine/cc188797.aspx
[58] (DIME): http://xml.coverpages.org/dime.html

Tables

Standard	Domain	Reference URL
(Dublin Core)	(Librarianship)	[23]
DIF	Science	[26]
Darwin Core	Biology	[28]
DDI	Social sciences, archiving	[31]
TEI	Social sciences, linguistics, humanities	[33]

Table 1: Common scientific metadata formats.

Standard	Jurisdiction	Reference URL
ISO 19115	International	[41]
CSDGM	United States	[42]
INSPIRE	European Union	[44]
GEMINI	United Kingdom	[45]

Table 2: Common geospatial metadata formats.

Attribute	Explanation
Name	Name of the keyword.
Data type	Type of the keyword, based on the Fortran 77 data types.
Dimensionality	Scalar or vector flag and, if appropriate size in each dimension.
Units	Units in which the value of the keyword is expressed.
External format	External format for displaying the keyword value.
Preferential display flag	An indication of the importance of displaying the keyword.
Comments	Comments describing the keyword.
Value	Value of the keyword.
Modification date	Date the keyword was last modified.

Table 3: Keyword attributes in the CURSA star catalogue manipulation system.